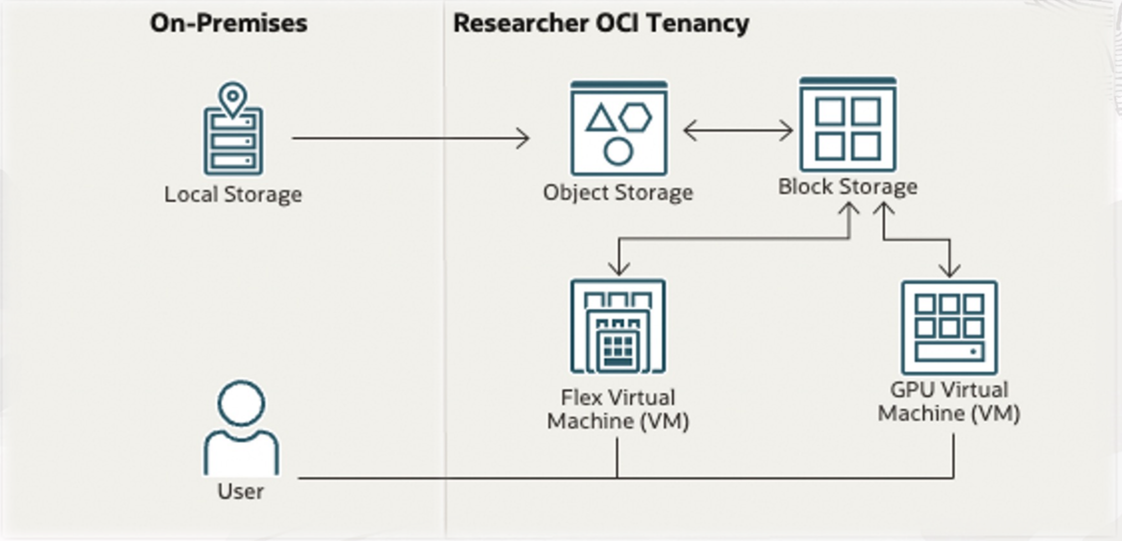
Technical Brief

University of Maryland and Oracle for Research explore discrimination's health effects

"Sticks and stones may break my bones, but words will never hurt me" goes the old playground adage. Unfortunately, research suggests that this simply isn't true, with racial discrimination being demonstrably linked to worsened health outcomes in ethnic minorities. University of Maryland are therefore developing artificial intelligence (AI) to identify discriminatory attitudes by location. This project from Maryland's School of Public Health is a representative case study in cloud-based, social media analytics through natural language processing (NLP) methods.

NLP-based applications and research are characterized by large datasets (e.g., 100's of GBs, or more) of unstructured data, and repeatable, memory-intensive computations. In this project, an NLP deep learning model is trained and used to classify (or infer) textual data. This technical brief provides an overview of some of the functional Oracle Cloud Infrastructure (OCI) components enabling this work. Similar endeavours developing or using NLP for analyzing/organizing large collections of documents, generating market insights, or other related NLP applications may benefit from this technical setup.



Cloud storage options for upload and analysis

Research projects using cloud processing against an existing dataset must first upload this data into the cloud. OCI offers different Cloud Storage options fitting various workload requirements and use cases. Storage options used in this project are Object Storage and Block Volumes.

Object Storage

Object Storage is frequently the first destination for uploaded data. It provides an unlimited capacity for unstructured data at a cost-efficient price point. Functionality offered by Object Storage includes versioning, replication, retention rules and pre-authenticated requests. Finally, it can be made easily accessible via the Console (e.g., browser-based interface), command line interface (CLI), various language SDKs, and a REST API. Because of this, Object Storage enables easy collaboration and a shared view of data amongst a research project team.

UMD's textual dataset is approximately 250 Gigabytes, which is characteristic of NLP projects (for another example, see details of the C4 dataset here) and is easily accommodated in Object Storage. In this project, the third-party, command line tool rclone was used to move data into Object Storage, however, other first- and third- party solutions are available.

Block Volumes

NLP training applications can benefit from the attached, performant read/write capabilities offered by Block Volumes. Volumes provide customizable storage sizes and performance levels. Attached volumes can be formatted and mounted to your compute instances and used in ways familiar to on-premise applications. They can be detached and attached later to quickly make data available to new instances.

For this project, Block Volumes allowed for fast access to data during GPU training and a means of persisting data and making it immediately available as the project moved from training to inference and analysis phases. Using the OCI Command Line Interface, data was copied from Object Storage and put into a Block-Volume-based file system. Results were similarly copied back to Object Storage for dissemination and downloading.

GPUs, virtual machines and flexible CPUs in the cloud

Model Training

As used in this project, and explored in detail in an OCI Blog post, GPU's provide a performant means of training NLP models (whether this be training from scratch or starting with a pre-trained Transformer model).

OCI offers multiple generations of NVIDIA GPUs across Virtual Machine (VM) and Bare Metal (BM) instances, including the *BM.GPU4.8* shape with A100's and the *VM.GPU3.1, VM.GPU3.2, VM.GPU3.4* and *BM.GPU3.8* shapes with V100s (more details here). Selecting a generation and quantity of GPUs allows for flexibility of problem size vs cost (GPU shapes charge per GPU/hour), and if needed multiple GPU instances may be used.

In this project, the much of the training was completed on a *VM.GPU3.2* shape, with 2 attached V100 GPUs (as well as 32GB GPU Memory, 12 CPU cores and 180GB CPU Memory). Following initial setup, a Custom Image was created and used to simplify setup of future instances.

Model Development (and Inference)

Not all work in a research project requires GPUs. Foregoing GPU shapes for 'flexible' CPU ones is a good way to match cloud spend to actual need. Flexible shapes allow hyper-customization across OCPUs (Oracle CPUs, see explanation here) and memory, and are well suited to tasks such as data exploration, initial code development, data movement and (sometimes) model inference. This research project benefitted from flexible instances of various sizes for data movement, code development, code optimization and inference.

Additional Services

Although not used in this project, OCI makes available additional PaaS and SaaS services that may be of functional interest to researchers working in this space. Oracle Data Science Service offers a suite of managed services for building, training, and deploying models, wrangling data and other related tasks. OCI AI Services provide prebuilt ML models that can be used for tasks including OCI Language, for text analysis at scale.

oracle.com/research