**ACCESS HPC Allocation Proposal**

**Elements: Data: Sustaining Modern Infrastructure For Political And Social Event Data**

**(NSF OAC-** **1931541)**

**November 1, 2022**

**PI: Patrick T. Brandt (Political Science, University of Texas, Dallas,** [**pbrandt@utdallas.edu**](mailto:pbrandt@utdallas.edu)**)**

**Co-PIs: Latifur Khan (Computer Science, University of Texas, Dallas,** [**lkhan@utdallas.edu**](mailto:lkhan@utdallas.edu)**), Vito D’Orazio (Political Science, West Virginia University,** [**vito.dorazio@mail.wvu.edu**](mailto:vito.dorazio@mail.wvu.edu)**) and Francisco Javier Osorio Zago (Political Science, University of Arizona,** [**josorio1@email.arizona.edu**](mailto:josorio1@email.arizona.edu)**)**

**Research Objectives**

This request is to support the NSF funded project (OAC-1931541) to collect news reports and generate event data from them. Event data are constructed from very large corpora of news reports. Given our need for additional resources in current allocation, the Jetstream2 HPC allows us to scale up to what is needed. It includes a mix of various nodes, great GPU capability, storage, and VM resources that meet the needs of our project.

**Storage, Performance, and resource choice**

All storage requirements are estimations based on the currently stored data in our UTD resources. Based on current allocation experiences, Table 1 lists the different Tasks and the nature of their computational and storage requirements.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tasks | Computation (CPU) | Computation (I/O) | Storage | Memory |
| Web Scraper | Moderate | High | High | Moderate |
| BERT Text Processing | Very High | High | Very High | Very High |

Table 1: Computational, Bandwidth, Storage and Memory requirements of each of the modules

|  |  |  |
| --- | --- | --- |
| **Storage Name** | **Current Size** | **Projected Size in a year** |
| Spanish Dataset & Training Models | 250 GB | 500 GB |
| Arabic Dataset & Training Models | 5 GB | 500 GB |
| **Total** | 255 GB | 1000 GB |

Table 2: Showing storage requirement for entire project.

**Justification of Service Unit Request**

Our need for Jetstream2 GPU and HPC capability is due to the lack of local resources available to us. When it comes to web scraping, local resources take a significant time to crawl websites due to the low number of CPU cores available to us. Most of the scraping is done on Google Collab which is not adequate for our work. When it comes to training BERT, we only have access to NVIDIA Tesla P100 GPU, which are a bottleneck for the scale of training we are planning to do. We’re having to reduce size of dataset or choose smaller models just to train on the P100. We need faster GPU and HPC resources to help facilitate our research.

Here is how we calculated SU as per Jetstream2 documentation available:

**Web scraping:** *VM cost per year= number of cores \* 24 \* 365.*

**Training (GPU Intensive):** *VM cost per year= (number of cores x 4 SUs/hour) \* 24 \* 365.*

Table 4 lists component wise required SUs. Table 5 points to prospective new module and related SUs required.

|  |  |  |  |
| --- | --- | --- | --- |
| Task | Number of Cores | Service Units/Day | Service Units/Year |
| Web Scraper | 16 | 384 | 140,160 |
| BERT Training Arabic | 8 | 768 | 280,320 |
| BERT Training Spanish |
| **Current Total** | 24 | 1152 | 420,480 |

**Table 4: Task and related SU usage summary**

From the data presented in Tables 4, our annual estimate for required SUs is 420480. Table 5 shows the requested e SUs and Storage estimated for 1 year.

|  |  |
| --- | --- |
| **Number of SUs per year** | 420,480 |
| **Storage requirement** | 1000 GB |

**Table 5: Project-wide required SU and storage requirements**