**Preparation for I590 - Big Data Open Source Software and Projects**

**Overview:**

In preparation for Dr. Fox’s Fall 2014 course on the High Performance Computing enhanced Apache Big Data Stack, this project will review the selected software, troubleshoot the installation process on FutureGrid resources, develop deployment scripts using Chef, and develop or adapt demos and tutorials to describe the functionality of the software packages and give simple hands-on examples of their practical use.

**Personnel:**

* Instructor: Dr. Geoffrey Fox
* Project Oversight: Scott McCaulay
* CS Master’s students: Anusha Gaddam, Karteek Pittala, Nikita Pandey
* REU student: Tatyana Matthews
* Other Staff: UITS Science Gateway Group

Each Master’s student will take responsibility for some number of software packages (X ~= total packages / # of students). They will be responsible for delivering scripts, demos, tutorials for each assigned package. When possible, packages will be assigned in meaningful combinations (e.g. Hadoop, HDFS, YARN).

The primary responsibility for our REU student will be testing scripts, demos and tutorials for clarity, accuracy and portability. This student may also take individual responsibility for 1 or more packages if time and workload allow.

The Project Coordinator will be responsible for training staff, monitoring and assuring progress, reviewing deliverables, and may also take individual responsibility for 1 or more packages if time and workload allow.

The Instructor will have ultimate responsibility for the list of included software packages, and the specific requirements for scripts, demos etc. with the understanding that the needs of the course may evolve over the duration of the project.

The specific contribution of the UITS Gateway group is not yet determined. They have offered their help.

**Selected Software Packages:**

Selected packages from Apache Big Data Stack:

|  |  |
| --- | --- |
| Airavata | Orchestration and Workflow |
| Hive | High Level (Integrated) Systems for Data Processing |
| Pig | High Level (Integrated) Systems for Data Processing |
| Mahout | Data Analytics: Machine Learning |
| Hadoop | Parallel Horizontally Scalable Data Processing:Batch |
| Giraph | Parallel Horizontally Scalable Data Processing:Graph |
| Spark | Inter-Process Communication:ABDS |
| Kafka | Inter-Process Communication:HPC |
| Memcached | Database:In-Memory Distributed |
| MySQL | Database:SQL |
| Hbase | Database:NoSQL:Column |
| MongoDB | Database:NoSQL:Document |
| YARN | Cluster Resource Management |

Other required software:

|  |  |
| --- | --- |
| Chef | DevOps/Cloud Deployment |
| Azure/Amazon | Commercial Clouds |
| OpenStack | IAAS System Manager |
| Cloudmesh | Management Framework |

Additional software packages still being considered:

|  |  |
| --- | --- |
| Mesos | Cluster Resource Management:ABDS |
| Slurm | Cluster Resource Management:HPC |
| Swift | File System:ABDS |
| HDFS | File System:ABDS |
| Lustre | File System:HPC |

**Deliverables:**

For each selected open source package, the following materials will be produced:

* Description
* Chef recipe
* Demo
* Tutorial

**Timeline:**

June 2, 2014 -> June 15, 2014

* Initiation, staffing, begin software evaluation and testing

June 16, 2014->July 11, 2014

* Finalize staffing, assign packages, begin developing materials

**Major Milestone -> July 11, 2014**

* Target date for first completed package from each master’s student

July 12, 2014-> August 17, 2014

* Agree on list of packages, deliver balance of documentation/demo packages, complete testing

August 18, 2014 -> August 24, 2014

* Any required follow-up, corrections, additional testing

**Major Milestone August 25, 2014**

* Fall classes begin