Predictive Analytics on Telemetry Data from HPE SimpliVity Customer Systems

Project Logistics:

Mentors: Jiri Schindler email: jiri.schindler@hpe.com; Peter Corbett email: peter.corbett@hpe.com;
Min-max team size: 3-5
Expected project hours per week (per team member): 6-8
Will the project be open source: no

Preferred Past Experience:

using Spark Nice to have
Java development Required
Scala development Desirable, but not required if prior Java experience
Familiarity with provisioning and automation tools (Ansible, Vagrant) Valuable
Mesosphere DC/OS or other Containers Infrastructure Valuable

Project Overview:

Background:
The ability to analyze and determine usage patterns across an entire install base, spot anomalies for a specific production system or predict future resource consumption for a customer deployment is an integral part of any enterprise system in a production environment. These analytics help administrators and IT personnel optimize operations and system providers (manufacturers) deliver more efficient and timely support.

The HPE SimpliVity HyperConverged Infrastructure is an on-premises solution that combines management of virtual machines (compute), storage with application and hardware lifecycle operations (disaster recovery, upgrades, system expansion). A SimpliVity solution consists of a collection of high-end commodity servers with local storage that are organized into clusters. Several clusters comprise a federation spanning different geographies. An entire federation uses a single management/operations console integrated into existing virtualization infrastructure solutions (e.g., VMWare vSphere). For more details on HPE SimpliVity see https://www.hpe.com/us/en/integrated-systems/simplivity.html).

Each system (node) deployed in a production environment sends periodically detailed configuration, telemetry, and performance data to HPE where it is processed and analyzed. The data is available for internal consumption by various organizations (product planning, architecture, performance engineering, support etc.). Since the acquisition of SimpliVity by

HPE in May 2016, we have embarked on a long-term project to revamp our platform for data processing and analysis of data we receive from tens of thousands of systems sold to-date to our customers. There are several scrum teams working on this effort, and in particular on the integration with existing HPE systems and databases tracking assets, customers and other information.

Project Specifics:
The focus of the project would be the development of new analyses and data visualizations of the configuration and telemetry data. Some of these may already exist as stand-alone purpose-built (one-of) tools or scripts developed by support engineers. One of the specific objectives would be to re-design and re-implement some of them in a scalable fashion and make them an integral component of our analytics platform: automatically pulling data from a common “data lake” and making the results available to everyone authorized within HPE for inspection.

Another project objective would be to work with the mentors (principal investigators) on developing new analyses such as unique data growth, compression and deduplication, resource consumption predictors, impact of snapshots and backups on overall storage consumption etc. The mentors will provide guidance in selecting suitable machine learning and data analysis algorithms and provide the necessary context. Therefore, background in machine learning or statistical analysis is not required.

The BU/NEU cloud computing course students would form a separate scrum team. They would be focusing on specific goals and tasks with minimal direct dependencies on other teams comprised of full time HPE employees that are part of the broader effort and focus on integration with other HPE systems including HPE InfoSight. The student team of students would follow our agile development processes with sprints and having their own backlog of stories. This separation will allow the students to make progress on their own, and cleanly delineate their work from that of the other teams. The students would follow (to the extent that is practical) our practices for developing production code (including writing unit tests and adopting continuous integration flow with automated deployment of a new version of the code onto a container infrastructure).

The overall effort dictates the selection of technologies, tools, and languages; the students would be expected to adopt them. We leverage open source technologies and tools readily available to everyone without special license, including, among others, Apache Spark, NiFi, Mesos (Mesosphere DC/OS), and Zeppelin. We also use CI tools like git, Jenkins, gradle, ansible etc. The expectation is that the artifacts (code and algorithms) developed by the student team would be directly

Some Technologies you will learn/use:

Operating as a scrum team that is part of a larger effort
Development of (private) cloud big data platform applications.
Data analysis and presentation of real-life customer systems data.
Continuous integration and automation practices necessary for production/large scale deployments
Interactions with senior technical executives and other experienced engieers.