Market Requirements Document

Feature Name: Objectivity/30 Intel ATK Integration

Version: 0.1 Date Submitted: 6/8/2015 Completed By: Brian Clark

Description of the Problem

Project ‘purple’ is bringing Objectivity products into the “Big Data” and “Fast Data” space leveraging a number of existing open source technologies. Intel have a “big data” analytics toolkit (IATK) for Apache Hadoop. We will partner with Intel and need to integrate ‘purple’ into the IATK.

This MRD will set out the requirements for the integration of ‘purple’ into the IATK.

Background

‘The revolution in “big data” is transforming industries and research, while spawning new solutions to a range of societal challenges. Big data strategies usually begin by capturing the high volumes of varied data using Apache* Hadoop*-based platforms that have massive scalability, cost effectiveness, and a vibrant open-source ecosystem. But once captured, achieving anticipated insights remains elusive. High demand data science expertise is scarce, exacerbated by added skills to program across a myriad of open- source tools and working through workflows inefficient for iteration and collaboration. Finally, tools used are often geared to answering known questions, with limited workable methods to easily find hidden signals in data patterns and connections.

The Intel Analytics Toolkit addresses these barriers to achieving value from big data and enables data scientists to achieve greater insights, more quickly, and with reduced complexity. A simpler programming environment lets data scientists focus on analytics instead of mastering the details of programming to Hadoop and the myriad of open source tools. Data scientists can orchestrate and easily iterate through the end-to-end analytics workflow in a single program, using a familiar programming language that executes analytics using fully scalable algorithms. Out of the box, the platform unifies entity based machine learning with an end-to-end graph processing pipeline including powerful algorithms for uncovering relationships hidden in big data. And the modular framework enables users or developers to extend and integrate new analytics functionality and algorithms.

By bringing simpler analytics programming and the full range of graph processing capabilities to the Hadoop* “data lake”, the Intel Analytics Toolkit is helping to democratize and accelerate big data powered solutions’.

Description of the Requested Feature

“The ATK is a platform that simplifies applying graph analytics and machine learning to big data for superior knowledge discovery and predictive modeling across a wide variety

of use cases and solutions. The ATK provides an analytics pipeline spanning feature engineering, graph construction, graph analytics, and machine learning using an extensible, modular framework. By unifying graph and entity-based machine learning, machine learning developers can incorporate an entitys nearby relationships to yield superior predictive models that better represent the contextual information in the data. All functionality operates at full scale, yet are accessed using a higher level Python data science programming abstraction to significantly ease the complexity of cluster computing and parallel processing. The platform is fully extensible through a plugin architecture that allows incorporating the full range of analytics and machine learning for any solution need in a unified workflow that frees the researchers from the overhead of understanding, integrating, and inefficiently iterating across a diversity of formats and interfaces."

There are 3 requirements:

1. The IATK supports data access through the use of Apache Spark Dataframes (Spark SQL). 'purple' is also implementing Dataframes to access data from Objectivity/DB. This implementation needs testing with the IATK.
2. The IATK the platform is fully extensible using a plugin architecture. This allows developers to expand the capabilities of the ATK for new problem solutions. Plugins are developed using a thin Scala wrapper, and the ATK framework automatically generates a Python presentation for those added functions. Plug-ins can be used for a range of purposes, such as developing custom algorithms for specialized data types, building custom transformations for commonly used functions to get higher performance than a UDF, or integrating other tools to further unify the workflow. We will need to implement the plugin to access Objectivity/DB.
3. The IATK provides a workflow capability through Python. Knime has been identified by Intel and Objectivity as a potential GUI to the workflow. There will be unit of work associated with integrating with Knime (see separated MRD for this integration).

Part of an existing feature or does it require another feature, if so, which one?

- New work.

How is this problem being solved now, and why isn't that acceptable?

- New work.

What languages must support this capability?

- Initially Java.

Which platforms must be supported?

- IATK runs on Linux (and Mac OS ?).

Do any competitors already have this feature?

- .

Customers who require this feature

- Some existing customers e.g. CGG, new customers in the ‘big data’ space.

Revenue at risk, or which could be won

- Could lead to more early adopters.

When is this required?

- Increments through ‘Purple’ MVP October 2015.

Additional Notes

1. **Implementation notes:**
    a. Need to prioritize what features are needed for each increment.

2. **Related Material**
   We will also need:
   Field Training.
   Quality Assurance.

3. **Software requirements**

4. **Hardware Requirements**
    a. Hadoop cluster