**Consortium for:**

**OPEN SOURCE ENTITY RESOLUTION ARCHITECTURE FRAMEWORK**

**05/12/2014**

**Alan Johnson, Qbase**

**Gregg Hanold, Qbase**

**James Conklin, CCRi**

**Adam Norris, Blue Canopy**

**Cameron Hunt, TRA Holdings**

**Shane Lawson, Ponte Technologies**

**Sponsored by Qbase, LLC**

**12018 Sunrise Valley Drive, Suite 300, Reston, VA 20191**

**Toll Free: 888 458 0345**

**Table of Contents**

Introduction 3

Executive Summary 3

Technical Demonstration 4

Infrastructure 6

Components used in the Technical Demonstration 6

**Table of Figures**

Figure 1 -OSERAF Operational View (Notional Multi-Level View) 4

Figure 2 – Reference Architecture 5

# Introduction

This white paper proposal outlines the establishment and technical demonstration of an Open Source Entity Resolution Architecture Framework (OSERAF) implemented with a specific set of prototypical use cases in mind. Specifically, this technical demonstration illustrates:

* The overarching principals and operational architecture of the OSERAF, with specific emphasis on its Entity Resolution Engine via a reference implementation of a select set of its technical components.
* The integration of multiple federated data sources, both structured and unstructured, with entity extraction and resolution, which provide a fused view into the data and its relationships based on a use case under analysis*.*

# Executive Summary

This white paper defines an Open Source Entity Resolution Architecture Framework (OSERAF) that provides for the fusion and analysis of content derived from multiple federated data sources that may be structured or unstructured, potentially accessed across multiple network domains.

A consortium of experts in the fields of Semantic Web, Natural Language Processing, System Security, Big Data Analytics, and Software and Hardware architectures as well as academia has been established that is uniquely positioned to apply methodologies, processes, and proven system security capabilities essential for solving the complex use cases surrounding data and information analysis vital to the commercial sector, our national security, and to the first response elements of our local, state and federal emergency responders. The OSERAF defined and proposed in this white paper demonstrates the capabilities of the consortium assembled to; conduct technology assessments to identify effective technologies; develop and integrate processes and technologies; test and evaluate systems; and transfer and integrate processes and technologies across multiple and diverse communities operating on multiple network domains.

The OSERAF establishes the operational architecture, processes, and systems that can be reproduced and adapted or subscribed to by commercial, local, state, and federal organizations involved in information sharing and analysis, civil infrastructure protection, emergency management, and Homeland Defense. Further, the OSERAF provides for establishing connectivity to such organizations as the private sector Information Sharing and Analysis Centers (ISACs), social networks, Joint Terrorism Task Forces (JTTFs), emergency management portals, law enforcement operations centers, state fusion centers, and military information operations centers. The operational view of the OSERAF is depicted in Figure 1.



Figure -OSERAF Operational View (Notional Multi-Level View)

# Technical Demonstration

The proposed technical demonstration of the Open Source Entity Resolution Architecture Framework (OSERAF) –implements, through Open Source Software (OSS), a subset of the total components necessary to support information dominance and address the current and future challenges identified with entity extraction, resolution, processing and analysis. The proposed implementation of the OSERAF framework will be configured in a hosted fashion via private Amazon EC2 compliant data centers. The reference implementation for the technical demonstration (Figure 2) consists of the following key elements:

* Collaboration Layer
* Visualization Layer
* Data Access Layer
* Common Access Layer
* Enrichment Engine
* Processing Engine
* Analysis Engine
* Data Stores (Graph)



Figure – Reference Architecture

The OSERAF that will be demonstrated provides the following key capabilities:

* High Performance, Multi-Modal Data Ingestion
* Structured
* Unstructured
* High Performance Data Processing, with a pluggable architecture
* Open API that supports a rich ecosystem of processing regimes
* Data Storage that is
* High Performance/High Volume
* Multi-faceted, flexible, Extensible
* Reliable and Secure (ACID)
* Interoperable
* Open
* High Performance Data Analysis, also with a pluggable architecture
* Open API that supports a rich ecosystem of processing regimes
* Highly appealing, user friendly Data Visualization and Collaboration environment that is/has
* Extremely compelling user experience
* Open, interoperable and pluggable
* Appealing to a broad range of users (novice to expert)
* Accessible in a variety of manners (UI, Command Line, multiple APIs)

## Infrastructure

The OSERAF is built in an Amazon VPC consisting of three Amazon EC2 micro server instances, each with 30G of EBS. The Domain Fusion Infrastructure micro-server is configured with the Collaboration Layer, Visualization Layer and Data Access Layer components. Given that the trusted agents necessary for a multi-domain implementation of the OSERAF require SE Linux extensions be applied in the AWS-EC2 Operating Systems, this is not a part of the technical demonstration proposed here. The multi-domain Infrastructure micro-server is configured with the protection layer, common access layer, processing engine, enrichment engine and analysis engine components and services. The data Infrastructure micro-server is configured with the relational, document and graph database components and their respective service.

## Components used in the Technical Demonstration

To demonstration the OSERAF, the following components will be employed in the associated layers, as described above:

* Collaboration Layer, Visualization Layer
  + Liferay Portal
  + Ikanow Visualization
* Data Access Layer, Common Access Layer
  + OrientDB Native Java, JDBC and Javascript API
* Enrichment Engine, Processing Engine
  + Ikanow Enrichment
  + Duke Link Analysis
  + Open System for Entity Resolution (OySTER)
* Analysis Engine
  + Ikanow Analytics
* Data Stores (Graph)
  + OrientDB

These particular capabilities are but a few of the many options that can be employed into the OSERAF; these have been chosen to effectively demonstration the technical approach in the quickest manner possible.

## Use Cases Considered

The Technical Demonstration of the OSERAF will focus on the following use cases thought to be of interest to the ISSO community:

Basic

1. Demonstration of Entity/relationship extraction/resolution from Iranian scientific documents (with emphasis on document related to potential Weapons of Mass Destruction (WMD)).
2. Interoperability demonstration between Ikanow and the balance of the OSERAF
3. Visual data representation, to include geospatial representations as well as link diagrams

Expanded:

1. Demonstration of various security models applicable to a property graph based storage paradigm, to include:
   1. Role Based Access Control
   2. Label Based Access Control (notional)
   3. Multi-Factor Access control (notional)

## Notional POA&M

A reference implementation of OSERAF is currently being established in the Amazon Cloud. This will greatly accelerate the implementation of this proposed technical demonstration, Key activities beyond that include:

**Plan Of Action & Milestones:**

Basic:

|  |  |  |
| --- | --- | --- |
| Activity | Milestone | Timeline |
| Identification of key data sources; development of property graph storage structure | Sources Identified, Structure Briefed | M1: Award + 30 days |
| Analysis, development, testing and technical demonstration of data access, extraction and loading mechanisms | Demonstration of data access, extraction and loading | M2: M1 + 45 days |
| Analysis, development, testing and technical demonstration of data comparators useful in performing entity resolution | Comparators Briefed; Demonstration of Entity Resolution against current corpus | M3: M1 +75 days (concurrent with M2) |
| Analysis, development, testing and technical demonstration of data visualization components useful in review of the enriched corpus produced by (2) and (3) operating on (1) | Demonstration of Visualization Components against current (enriched) corpus | M4: M1 + 90 days  (concurrent with M2 and M3) |

Expanded:

|  |  |  |
| --- | --- | --- |
| Activity | Milestone | Timeline |
| Establishment of appropriate identity and access management components | Demonstration of identity and access management components | M5: M1 + 45 days |
| Analysis, development, testing and technical demonstration of select security models to the property graph structures | Demonstration of access control based on select security models | M6: M5 + 45 days |

## Rough Order of Magnitude Price

To support the technical demonstration outlined here, the following Labor and Material costs are anticipated (based on blended rate of $150/hour):

