TRILLION TECHNOLOGY SOLUTIONS

SMART DATA ANALYTICS

Cyber-Attack Detection and Fraud Prevention Using Data Analytics

Challenge

Traditional approaches to cyber-attack and associated fraud rely on single platform detection of known incident signatures. This challenge is common across multiple sectors including financial, healthcare, government agencies such as intelligence, defense, homeland security and emergency management, as well as e-commerce. Damaging security breaches have been reported to many organizations with many more unreported or undetected. Clearly cybercriminals, rogue nations, terrorists and hackers are waging war with increasingly sophisticated approaches to steal and sabotage nations, businesses and ordinary citizens. Today's fast Internet and high-performance computing provides them with means to flood the environment looking for vulnerabilities at rates too high for conventional attack detection approaches.

Trillion's Solution

Cyber-attack detection and fraud prevention is inherently a data analytics problem limiting current detection approaches. It requires analyzing multiple types of data including streaming, structured and unstructured data. Trillion's Smart Data Analytics Solution (Figure 1) applies an architecture that leverages multiple Big Data eco-system components offering a variety of heuristics, probabilistic, and signal detection analytics to support this complex detection problem. The premise of our approach is large scale, sophisticated cyber-attack challenges

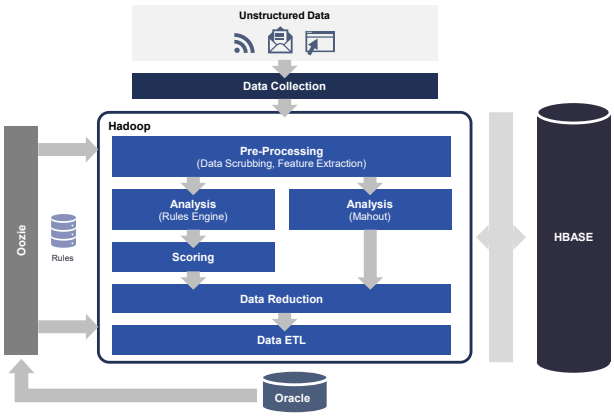

Figure 1: Smart Data Analytics Solution

requires large scale, sophisticated intelligent detection solutions which are customizable, modular and scalable.

Trillion's strategy extends the traditional cyber-attack and fraud detection mechanism of monitoring system activities for signs of ongoing attacks with massively parallel processing. Raw unstructured data feeds are decomposed into parallel efforts to extract or scrape data elements needed to support feature extraction. These features are further decomposed into parallel efforts applying various normalizing, sanitizing, heuristic, pattern recognition, and probabilistic machine learning analyzers providing a collection of detection scores. These scores are reduced to provide classification and categorizing results to be extracted for review and action. This data flow is captured within a Hadoop ecosystem providing the scale needed to process 10 Terabytes of raw data into detection results within 15 minutes. This architecture provides flexibility by combining rules with machine learning algorithms.

The platform generates complex models providing predictive analytics for cyber-attacks before they happen. The models are continuously trained. Models that achieve pre-determined thresholds are fed into the rules engines to increase the probability of the final models that are applied to the target environment. Historical trends and real-time data feeds allow the system to generate operational recommendations with very high confidence percentages.

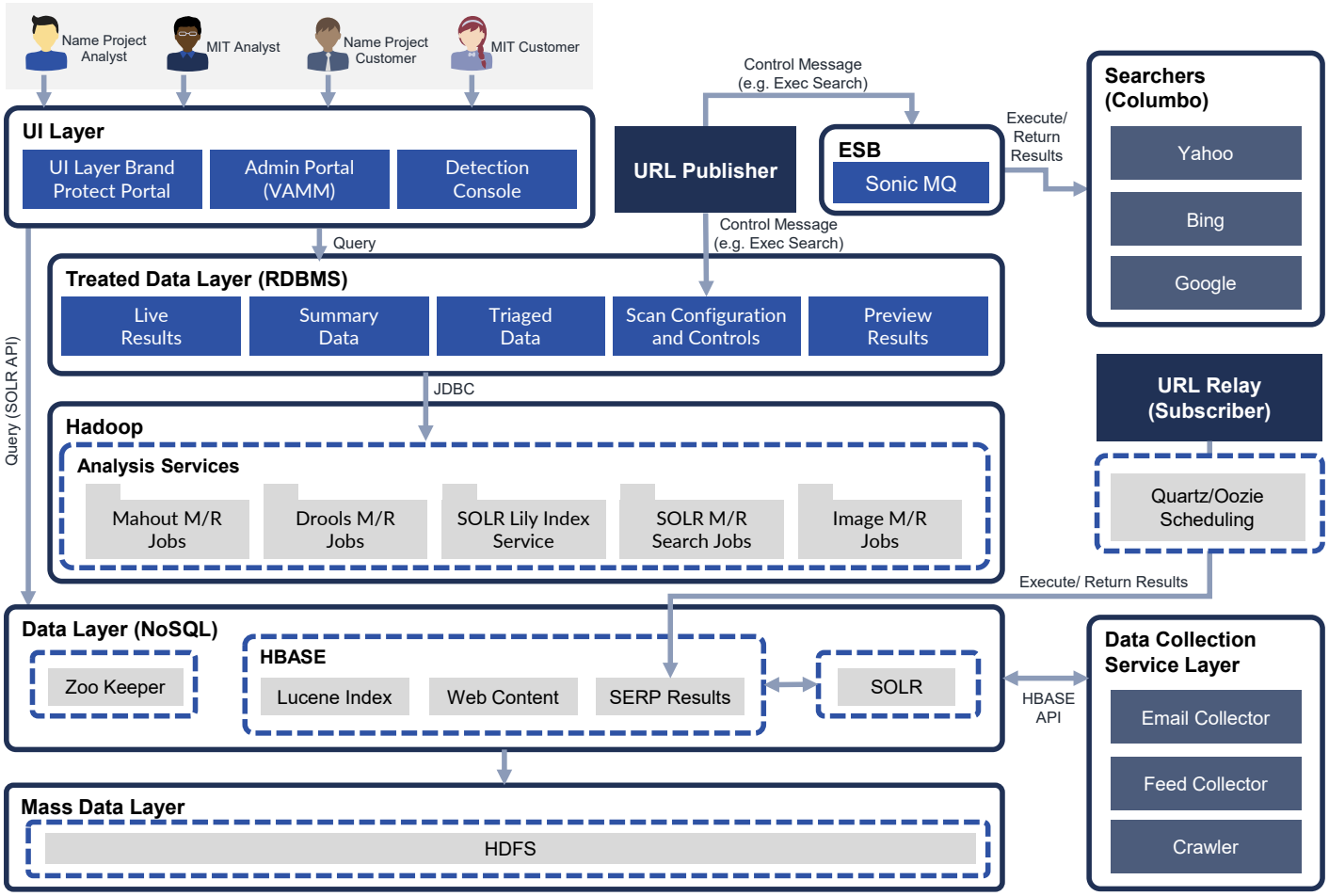

Figure 2: Anti-Phishing Platform

The Anti-Phishing Platform (Figure 2) analyzes emails and webpages for potential phishing attacks on customers. A series of ETL and analysis components were designed and implemented to filter, analyze and identify content that could potentially be phishing attacks. The platform consists of data collectors, crawlers, machine learning components and analyzers. Analytics include Mahout Classification machine learning capabilities for data mining and logistic regression classification applications. Several other analytic methods are applied in parallel including Drools rules-based pattern recognition and heuristic inference engines. SOLR based search engines search for rules based key words while image matching algorithms are used for image/logo comparisons. Integration with queueing frameworks provides seamless integration across various components. The system seamlessly processes data in batch mode as well as real time streaming input.

Scalable and Extensible

The inherent nature of the architecture is scalable by expanding and creating additional parallel processing across multiple components. With minimal adapter changes, components are easily swapped with equivalent technologies due to the modular architecture of the solution. The core architecture can be used across multiple verticals such as emergency management, national security, intelligence, financial, healthcare, and retail e-commerce industries. There is a wealth of historical and current real time data available across all these verticals for which this solution can easily be customized and extended for smart analytics and visualization.

An extended version of our solution incorporates Apache Spark and Apache Zeppelin for faster analytics and visualization. We are developing a proof-of-concept with Apache Apex and Apache Flink as alternatives for real-time streaming and distributed processing. Our modular extensible architecture easily allows these permutations and combinations for a targeted solution based upon the applicable vertical, government agencies or commercial industry.

