Applying High-speed Pattern Recognition to Generate Queryable Semantics from Big Data

Chuck Rehberg

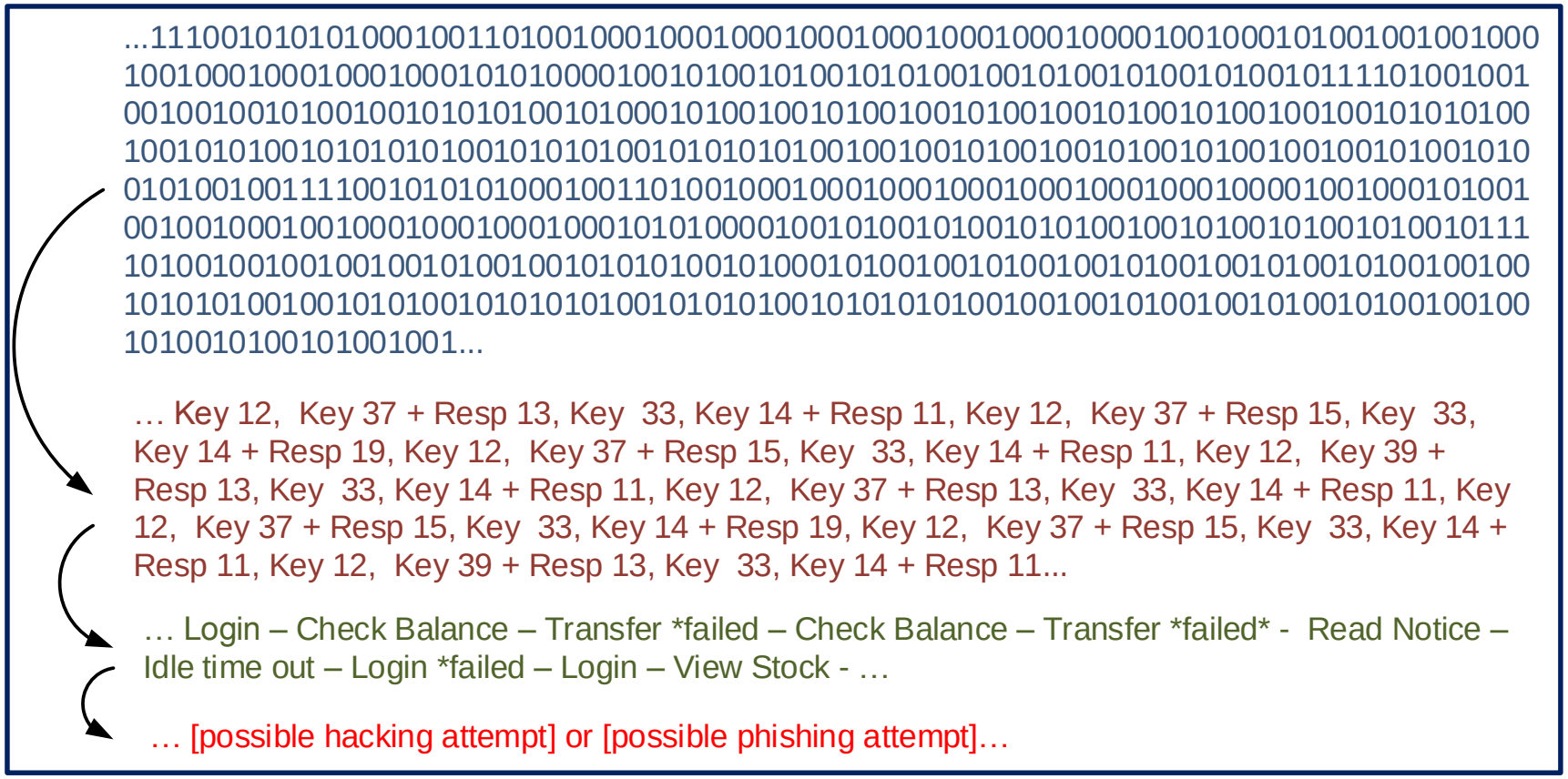
CTO Trigent Software

Chief Scientist at Semantic Insights

October 2014

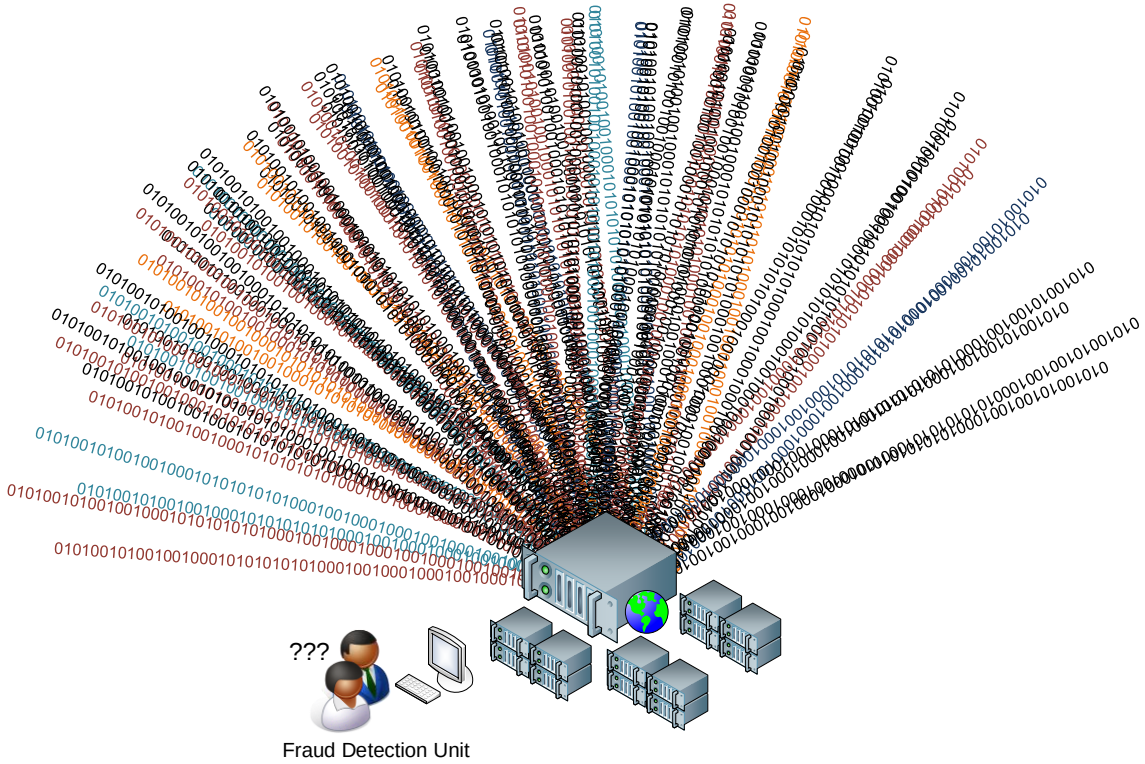
The Problem: Using Raw Data for Abstract Reasoning

In some cases, where big data is collected in real-time, the data being collected is at a level of abstraction that is too low for domain-knowledgeable investigators to directly query.



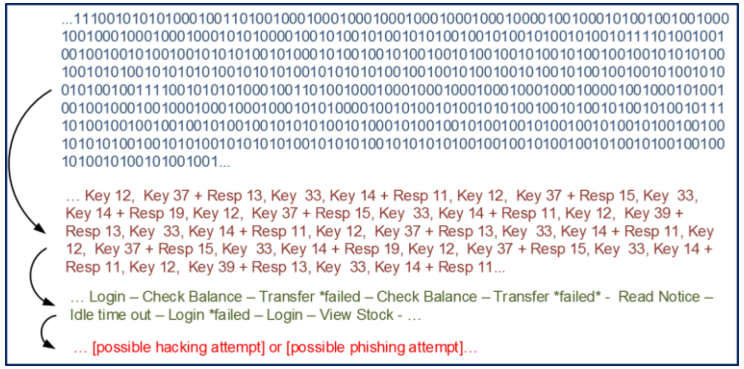
Example: Data at the wrong level of Abstraction

1. Suppose you have a data stream where a massive amount of real-time data = “individual key clicks + responses” tagged to individual interactive sessions.
2. Suppose domain-knowledgeable investigators are interested in a group of nearly simultaneous transactions from multiple distributed interactive sessions involving any one of a set of specific data items.
3. Suppose the “key clicks + responses” tagged to each interactive session are interspersed with a wide variety of other (non-relevant) “key clicks + responses” within the same interactive session.
4. Suppose the domain-knowledgeable investigators wish to conduct their investigation with ad-hoc queries that are not in terms of “relative sequences” of “key clicks + responses”. Rather, they wish to make queries in terms of high level transactions.
5. And finally, suppose you have more than one massive interspersed data stream with similar and/or related information, which taken together are used to infer the high level transactions referenced in the queries of the domain-knowledgeable investigators.



A Semantic Translation Solution

- The raw data represents the lowest level of abstraction available. Often it is the only level of abstraction available.
- A high-speed complex pattern recognizer can be employed to sift through and correlate, recognize and generate aggregate data representing a higher level of abstraction. Multiple possible interpretations can be maintained.
- Additional high-speed complex pattern recognizers can operate on this new higher level data to generate new aggregate data representing an even higher level of abstraction. This can be continued to allow flexible changes to each data transformation level. Again, multiple possible interpretations can be maintained.
- *The goal is to produce data representing higher levels of abstractions which constitute semantic representations required by the domain-knowledgeable investigators.*



Two Key Translation Challenges must be met

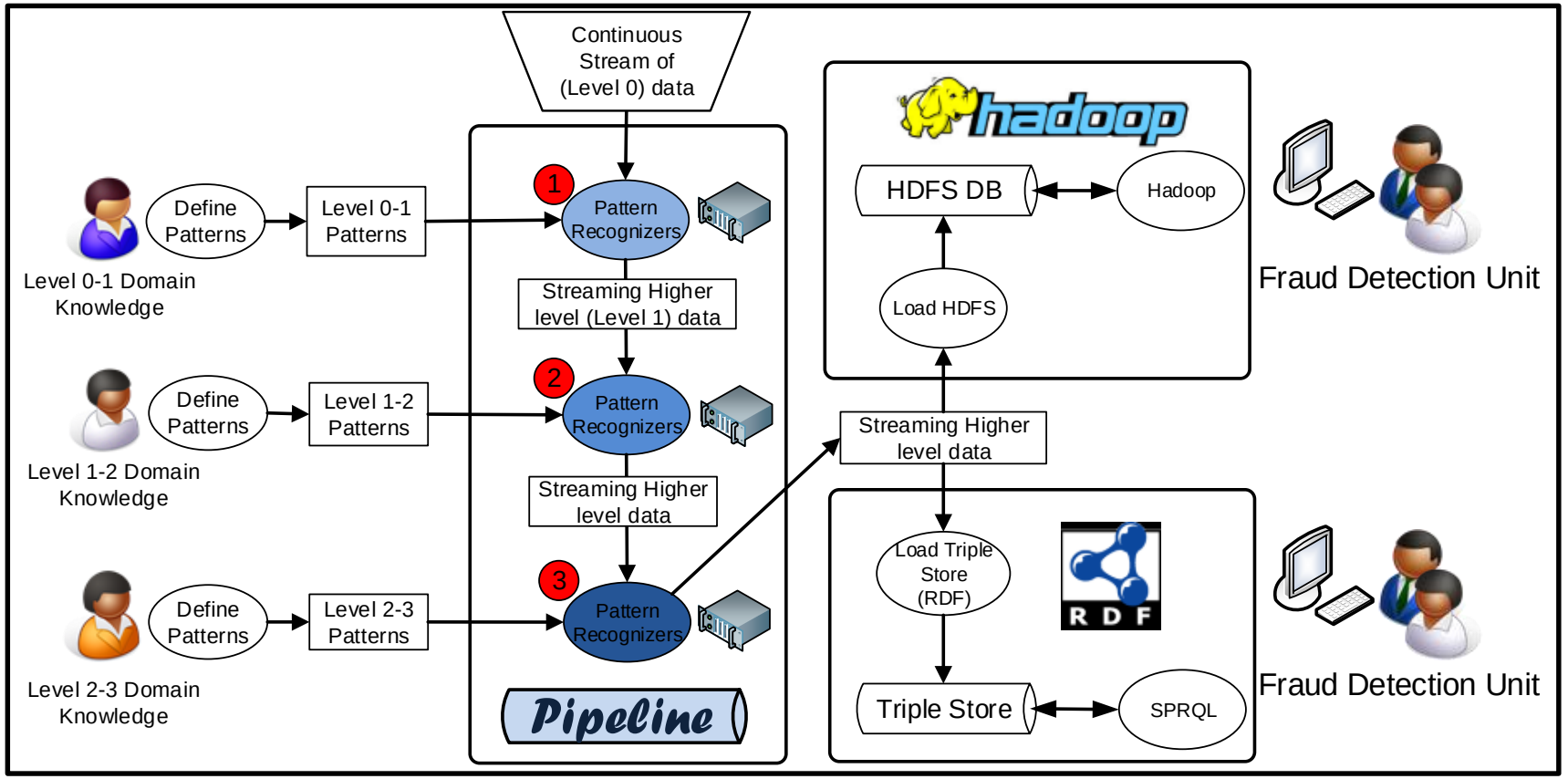
1. One key challenge is to flexibly transform the raw data into data that domain-knowledgeable investigators can directly query.
   - The transformation may require considering at a vast amount of data in sequence before a proper transformation can be made.
   - Multiple possible interpretations of the data may need to be maintained.
   - Individual transformations should be allowed to evolve over time to become more accurate
   - The data may contain information that does not contribute (or even detracts from) the semantic interpretation of the data. The data of interest may be surrounded by semantic “noise”.
2. For time-sensitive data, you will need to transform the data in real-time as it arrives.

Real-time Semantic Transformation

- Our High-speed complex pattern recognizer is based on the Fast Rule Selection Engine (FRSE). US Patent #7,433,858
  - We have made additional modifications to the FRSE Engine to efficiently handle time-sequenced pattern matching.
  - Patterns can match adjacent tokens or span intervening non-matching tokens (noise).
  - The number of tokens to be scanned at a time can be pre-specified or adjust dynamically as needed.
- Multiple simultaneous levels of transformation in real-time is accomplished by pipelining.
  - Where each level hands off to the next level for further processing while continuing to process the new input.

Transforming Streamed Data in Real-Time: A Multi-layered Pipelined Approach

1. The continuous stream of data is transformed (1) into objects recognized in another transformation (2) process which in turn produces objects that are sent to a third transformation (3) process, and so forth.
2. This can continue an arbitrary number of times with each transformation running independently in parallel.
3. This is well suited to problems requiring a cascade of information toward increasing levels of abstraction.



Specifying Transformation Patterns Flexibly

Transformation Patterns processed by the FRSE Engine are powerful and simple to define:

1. ***Transformation Patterns*** are sequences of ***Token Expressions*** (i.e. expressions that match ***Match Tokens***)
2. ***Transformation Patterns*** also produce a ***Pattern Result***
3. ***Pattern Result*** creates a new sequence of one or more ***Result Tokens*** in the output data
4. ***Match Tokens*** are any data item that can be matched in the input data
5. ***Result Tokens*** are created by ***Transformation Patterns***
6. ***Token Expressions*** state what must be true for a ***Match Token*** to be correctly identified
7. ***Result Tokens*** can act like ***Match Tokens*** in subsequent ***Transformation Patterns***

Specifically:

```
<Transformation Pattern> ::= <Token Expression>(1,n) --> [<directive>] <Result Token>(1,n)
<Token Expression>  ::= [<Match Operator>] <Match Token>    /* default Match Operator is '=' */
<Match Operator>    ::= { = | <> | > | < | <= | => }
<Directive>         ::= { ADD | REPLACE | RESULT }          /* ADD/REPLACE modifies input; RESULT sends to output */
<Match Token>       ::= <Token Set ID> | token(1,n)
<Token Set ID>      ::= token(1,n)                          /* Token Set ID acts like a synonym */
<Result Token>      ::= token
```

Summary

- Big Data, representing real world events, often needs to be filtered and transformed into semantically meaningful data before domain-knowledgeable analysts can perform queries.
- Today, interactive data visualization techniques are often used to display *data aggregations and relationships* to help domain-knowledgeable analysts preform generalized “visual queries”.
- However, data visualization can be particularly unrevealing when data aggregation requires recognizing content-dependent patterns in a field of physical and semantic noise.
- High-speed Pattern Recognizers are capable of filtering and transforming real-time Big Data for immediate high-level evaluation.
- For more information, please visit us at www.semanticinsights.com.

Chuck Rehberg

As CTO at Trigent Software and Chief Scientist at Semantic Insights, Chuck Rehberg has developed patented high performance rules engine technology and advanced natural language understanding technologies that empower a new generation of semantic research solutions.

Chuck has more than thirty years in the high-tech industry, developing leading-edge solutions in the areas of Artificial Intelligence, Semantic Technologies, analysis and large –scale configuration software.

SEMANTIC INSIGHTS™

A Division of Trigent Software, Inc.