A CRIME INVESTIGATION TOOL BASED ON EVENT PATTERN DETECTION :

by

Osbert Osamai

Reg. No: 2010/HD18/1259U

BITC (KYU), DCS (KYU)

Department of Networks

School of Computing and Informatics Technology, Makerere University

E-mail: omeja4ever@gmail.com,oosbert@trobank.com

Tel: +256712738530

A Project Report Submitted to the School of Graduate Studies in Partial Fulfillment for the Award of Master of Science in Data Communication and Software Engineering Degree of Makerere University.

OPTION: Mobile Computing

Supervisor

Kanagwa Benjamin (PhD)

Signature ———————————

Date ———————————

Department of Networks

School of Computing and Information Technology, Makerere University

Email: bkanagwa@gmail.com Tel: +256-712495020

March 2013

Contents

ACRONYMS **i**

1 Introduction **1**

1.1 Background 1

1.2 Statement of the problem 2

1.3 Objectives 2

1.3.1 General Objective 2

1.3.2 Specific Objectives 2

1.4 Scope 3

1.5 Significance of Project 3

2 Literature Review **4**

2.1 Current Technologies in Crime Investigation Systems 4

2.2 Related work in Uganda 5

2.3 Complex Event Processing 5

2.4 Event Processing Languages 8

3 Methodology **12**

3.1 Requirement Gathering 12

3.1.1 Interviews 12

3.1.2 Existing Literature 12

3.1.3 Questionnaires 13

3.2 Requirements Review 13

4 The Crime Investigation Tool 14

4.1 System Design 14

4.1.1 Data flow diagram 14

4.2 System Implementation 15

4.2.1 System 3-Tier Architecture 16

4.3 Crime Patterns Considered 17

4.3.1 Patterns Related to Bank Account Activity 17

LIST OF ACRONYMS

CEP	Complex Event Processing
SNA	Social Network Analysis
GIS	Geographical Information Systems
SPP	Spatial Point Patterns
SQL	Structured Query Language
SOA	Service Oriented Architecture
JDBC	Java Database Connectivity
ODBC	Open Database Connectivity
SOA	Service Oriented Architecture
XML	Manipulation Language
CSV	Comma Separated Values
POJO	Plain Old Java Object
SDLC	System Development Life Cycle

1 Introduction

1.1 Background

Crime investigation world over is considered a difficult and laborious process that heavily relies on efficient crime analysis to ensure accurate and timely conclusions. This is not helped by the fact that law enforcement agencies deploy manual investigation processes and computerized systems that cannot quickly identify complex crime patterns.

In Uganda, the situation is no different with these agencies facing massive delays in crime solving and increasing numbers of sophisticated crimes, most of them of the organized. Every tens of thousands of cases are carried forward to the following year, uncompleted. As the usual circle of crime would dictate, fresh cases are reported every day, and, gradually, older cases left uncompleted lose the urgency they initially generated and, inadvertently, they die a natural death [11].

To avert this problem there is need to deploy sophisticated tools, technologies and resources that can enable crime investigators quickly reach reasonable conclusions by identifying patterns of behavior in criminals. In so doing, not only will crimes be investigated faster but some future crimes may be prevented based on recurring patterns identified. However, intelligence and law enforcement agencies are often faced with the dilemma of having too much data, which in effect makes too little value. On one hand, they have large volumes of raw data collected from multiple sources: phone records, bank accounts and transactions, vehicle sales and registration records, and surveillance reports. On the other hand, they lack sophisticated network analysis tools and techniques to utilize the data effectively and efficiently [15].

In this project the phenomenon of Complex Event Processing (CEP) is used to detect patterns in crime related events using the Esper engine for high throughput and performance. CEP analyses low level events to produce a single complex event and has been successfully used in Stock Trading, Network Analysis and other areas.

1.2 Statement of the problem

Due to the manual nature of crime investigations in Uganda, case backlog remains a critical issue leading to delayed justice, unpunished criminals and of course tainting the image of the agencies. These manual methods often lead to fatigue, poor statistical analysis and the inability to solve crimes through pattern detection and analysis.

Criminals have become very intelligent due to the advancement of technology and therefore they conduct crimes in an untraceable manner. Majority of those crimes evolve in a long period of time making them even more difficult to predict. Therefore the rate of organized crime is on the rise most of which are orchestrated in vast geographical areas using these complex techniques.

Therefore, manual techniques of analyzing such data with a vast variation have resulted in lower productivity and ineffective utilization of manpower [3].There is need to develop a tool to quicken the investigation process by accurately guiding investigators in evidence analysis and also use recurrent patterns to prevent future crimes.

1.3 Objectives

1.3.1 General Objective

The objective of this project is to design, develop and deploy a crime investigation tool to guide investigators by detecting crime patterns and monitoring criminal activities. The tool will operate by filtering criminal activities as events and detecting predetermined patterns. This will not only reduce on time spent during investigations but also prevent some crimes from occurring by discovering recurring patterns.

1.3.2 Specific Objectives

Specific objectives include:-

1. Design and build a crime investigation tool based on complex events with a processing engine and web-based visual interface.

2. Thoroughly validate and test the solution.

3. Deploy the solution in a live working environment.

1.4 Scope

The project covers general aspects of using complex event processing in applications but the implemented tool processes data from an existing information system and only focuses on crime patterns inherent in financial embezzlement related crimes. Document processing and forensics related actions are not covered in this project.

1.5 Significance of Project

The benefits of the project apply to the Police and the general public. These include:-

1. Improve efficiency by reducing manual operations and concentrating on analysis and prediction efforts hence also reducing on the time spent during investigations.

2. Reduce on crime rates by preventing their occurrence through discovery of uniform patterns.

2 Literature Review

This section provides a general literature review of major data mining techniques used in existing crime investigation systems and Complex Event Processing as the preferred technology for this project.

2.1 Current Technologies in Crime Investigation Systems

Existing crime investigation systems tend to vary in terms of their overall capabilities and technical operation. In one study [2]the existence of prominent criminal investigation software like HOLMES2, BRAINS and Analysts' Notebook which are used by criminal analysts in the United Kingdom and Holland was acknowledged. This category can be classified as early generation systems that mainly focused on analyzing evidence separately without linking multiple sets of evidence in order to solve crimes. They were also not designed to communicate with existing case management systems to facilitate data exchanges.

The Second generation systems around the world relied heavily on data mining techniques to query large datasets for meaningful patterns in order to help investigators solve crimes. These methods however fall short in terms of supporting decision making largely due to poor processing speeds and querying algorithms.

Link Analysis is a technique used in data mining. These tools have for long been used by law enforcement agencies to identify, analyze and visualize relationships between crime entities. In a study [5] , it is revealed that through association paths linking suspects and victims in crime, link analysis discovers information about motives and hence provides investigative leads.

Another technique used in crime pattern detection involves several data mining steps like hotspot detection, crime clock, crime comparison and crime pattern visualization. Numerous algorithms are used to relate multiple crime scenes, represent a number of crimes scenes on a daily basis, compare different crimes to estimate growth rates and visualize the changes in crime occurrence frequencies [3].

A study on crime network analysis [15] suggests that law enforcement agencies need to deploy reliable data and sophisticated tools as critical tools in the discovering useful patterns in data. They introduce a data mining techniques called Social Network Analysis (SNA) which is used to discover hidden patterns in large volumes of crime related data. An approach of SNA referred to as black modeling is used in criminal networks to reveal associations between subgroups based on a link density measure. Discovery of new structural patterns during this process can enable prevention of crimes and also modify conventional view of certain crimes by investigators.

2.2 Related work in Uganda

In Uganda, some studies[16] have been conducted in the area of crime investigation. one research discusses a model for forensic investigations that performs detection of incidents through system monitoring and performs data analysis to unearth the crime scene, suspect and how the crime was perpetrated. The study proposed a new model based on five iterative phases that were meant to strengthen the crime detection and analysis process.

Another study [17] on crime prevention suggested a combined application of data mining techniques alongside GIS (Geographical Information Systems) to discover crime data in disorganized settings like Uganda. Spatial point patterns (SPP) based on coordinates of events such as locations of crime incidences and the time of occurrence are used.

2.3 Complex Event Processing

A Complex event is an event that abstracts or aggregates simple (or member) events [7] . Simple and complex events are normally represented in linear ordered sequences called Event Streams. These streams are usually bound by time intervals and may contain different types of events.

Complex Event Processing (CEP) is defined as the process of detecting complex events using continuously incoming events on a lower abstraction level[4]. This study justifies the need for CEP given the fact that single events on their own may not be sufficient in determining certain patterns. CEP therefore provides a platform for a combination of events to be processed and analyzed.

CEP is a foundational technology for detecting and managing the events that happen in event driven enterprises. It is a collection of methods, tools and techniques applied in processing events as they happen. In order to achieve a lot from CEP, happenings of events in enterprises need to be well understood. This can be achieved by organizing events into structures or hierarchies, identifying relationships among events (causal, time, aggregation) and organizing events in different views from different personnel. In CEP, higher- level knowledge is derived from lower-level events which are a combination of various occurrences. CEP can be viewed in two types, the first one involving specification of complex events as patterns and detecting them effectively, whereas the other type involves detecting new patterns as complex events. In the first case, event query languages offer convenient means to specify complex events and detect them efficiently. In the second case, machine learning and data mining methods are applied to event streams.

Detection of complex events is, of course, no an end in itself; an event-driven information system should react automatically and adequately to detected events. Typical reactions include notifications (e.g., to another system or a human user), simple actions (e.g., buy stocks, activate fire extinguishing installation), or interaction with business processes (e.g., initiation of a new process, cancellation or modification of a running process).

A study about the history of CEP [18] traces its roots to university and company research groups in the late 90's which were involved in the areas of active databases, event driven simulation, networking and event processing in middleware. This explains partly why the CEP query languages are based on SQL syntax having been influenced by research on active databases. CEP products at that point did not generate much interest until the late 2000's when CEP was deployed as add-ons on SOA architectures and ESBs.

Databases are distinct from event queries used in CEP because of the latter's ability to continuously detect events as they happen rather than just acting on stored datasets. Event processing languages need to enable the possibility of joining several individual events together, so that their combined occurrences over time yield a complex event and complex events must contain the element of time, to track times when events occur.one study introduced the need for revision of events in cases of erroneous data. In practice, there are a number of reasons requiring revisions in event stream processing. For example, an event was reported by mistake, but did not happen in reality (and the mistake was realized later); an event happened, but it was not reported (due to failure of either a

sensor, or failure of the event transmission system); or an event was triggered and later revoked due to the transaction failure. Also very often streaming data sources contend with noise (e.g., financial data feeds, Web streaming data, updates etc.) resulting in erroneous inputs and, therefore, erroneous complex event results.

Through Complex Event Processing (CEP), companies and organizations can manage processes in close to real-time. It is however noted that due to the complexity of generic event processing frameworks offered by the industry, the configuration and setup of CEP applications are left to external experts who are more knowledgeable in complex event logic. A CEP application retrieves events for all noteworthy incidents in the business environment. In various parts of the application, event-pattern rules are applied on the incoming event stream to detect relevant patterns, e.g., an uptrend in application errors or execution delays. In response to such patterns, the CEP engine proactively intervenes in the business environment, e.g., by temporarily allocating additional resources, throttling uncritical business tasks or notifying system administrators [18].

Pattern matching is a key feature of all CEP technologies which involves finding subsets of data matching a given pattern and also relationships between those subsets. A study about CEP under uncertainty explains the role pattern matching plays in allowing users to look beyond individual events and find specific collection sets.

Solution templates are proposed to perform data mining procedures on historical data to identify event patterns besides real-time monitoring. The central concept of any template's event processing infrastructure is the event processing map, a predefined orchestration of event adapters and event services. Event adapters may be considered the actual interface to the underlying source system: Depending on their implementation, event adapters translate real-world actions (such as a user actually placing a bet in an online gambling platform) into event representations of a certain event type, and vice versa. Event services receive events from event adapters or other event services, process them based on implementation of specific logic, and respond back to the map.Detecting events is based on some considerations like, some events sharing time elements, the order of events, time bounds within events and detection of events of long time lags. To gain insight into processes there is need to include the following components in CEP application, Facilities (graphical or textual) for precise description of complex patterns of events, Scalable performance, modular rules engines to detect complex patterns of events, Facilities for defining and composing event pattern

triggered rules for pattern abstraction.

2.4 Event Processing Languages

An analysis of Event Processing Languages [4] revealed the need to shift from using general-purpose languages like C, Java, C++ e.t.c for CEP applications due to low-level complexities. Using such languages along with complexities like data structures and algorithms can only complicate the development process. The study provided a detailed analysis of existing CEP programming Languages and platforms based on their expressivity and integration capabilities. Expressivity is measured by one of the following abilities, filtering streams by event type, processing a subset of events (windows), data extraction and aggregation of data over events, performing conjunctions and disjunctions, show temporal relations between events, showing causality of events, negation and counting of events, event instance selection and consumption to prevent reuse of events in pattern detection and integration of event data and non-event data (data from outside). Languages are also grouped into the categories of data stream query languages, composition-operator-based languages, production rule languages and logical formulas.

STREAM (Stanford Stream data Manager) is a language whose focus was to develop methods to manage and query data in data streams and was a result of a research project at Stanford University. The project also produced a CEP engine called STREAM and an Event Query Language called Continuous Query Language to query events. STREAM was a basis for other data stream languages like Esper and its querying syntax resembles SQL very strongly.

Borealis is a CEP engine developed at Brandeis University and MIT that uses a ”boxes and arrows” approach. Queries are described graphically with queries as boxes and streams as arrows connecting boxes. The approach was first used in an earlier engine, Aurora. The main difference between Borealis and stream languages is the focus on query evaluation that Borealis offers resulting in less abstract queries than STREAM.

Active Middleware Technology (AMiT) enables IBM middleware to become event-based. This technology is implemented in several products, most notably extending Web- Sphere Broker with CEP capabilities. As WebSphere is a commercial product, it is not freely available (requires regis-

tration .Basic events are declared with their attributes in event tags. Lifespans are windows defined by two events, an initiator and a terminator event. Lifespan types are therefore declared by referencing start and end event types. Whenever an event matching the initiator speci?cation is detected, a new lifespan of this type is opened, and when an event matching the terminator specification is detected, the lifespan is closed. Complex events are called situations. A situation consists of at least one data attribute (it has to carry at least one kind of information), exactly one operator, and a lifespan type. Situations are only tried to be detected in lifespans of its type. A lifespan may be referenced by multiple situations

RuleCore is a CEP engine developed by Analog Software, building on research at the University of Skyde. As the name suggests, rules are the central concept of ruleCore. The ruleCore engine processes events using ECA (Event-Condition-Action) rules that consist of three parts: for every event (basic or complex), check a condition; if it is true, execute the action. ruleCore has two implementations; an open source variant called ruleCore, released under the terms of the GPL; as well as a commercial version called ruleCore CEP Server. RuleCore uses so-called detector trees for event detection. Leaf detector nodes (detector nodes without children) detect single events (they pick up events of their type). They are inactive until an event of their type is delivered to the rule (usually by entering the system, although exceptions are mentioned in the next paragraph), after which point they are always active. To detect complex events, a detector tree is built: the leaves detect simple events, and inner nodes detect complex events depending on whether its children detected events.

SASE+ is a CEP system developed at the University of Massachusetts, Amherst. It is an extension of the older SASE system. The system is designed for event streams with many events per time unit and also queries using large time windows, creating new issues regarding efficient query execution. The project's purpose is to devise techniques for high-performance querying of event streams, using a declarative, composition-operator-based language. Although SASE+ is an agile language and concentrates only on pattern matching on streaming data, the pattern matching properties of SASE+ can be used in more general contexts.

Esper is an open-source CEP engine, developed by EsperTech Inc. and volunteers, released under the GNU General Public License (GPL v2). As stated on the official web site, it is designed for CEP and Event Stream Processing (ESP).There are two implementations of Esper, Esper for

Java and NEsper for .NET. Both supply an API to access the engine features, such as deploying queries, sending events into the engine and retrieving events out of the engine, in their respective language. Events are objects in their respective language; for Esper, events can be instances of java.util.Map,org.w3c.dom.Node (Java representations of XML documents), or other Java objects. Regardless of the implementation language, queries are stated in a SQL-like language called Event Processing Language (EPL).

Cayuga is a research CEP engine developed at Cornell University. It sets itself apart from other engines in that it deliberately sacrifices expressivity for performance, targeting applications running large numbers of queries. It is free software, available under the terms of the BSD license. Cayuga uses an Event Query Language called Cayuga Event Language (CEL). While its syntax resembles SQL, like many data stream languages, it also offers patterns, although using a different approach compared to Esper, inspired by regular expressions.

Drools, also known as JBoss rules, is a production-quality business rule management system, including a production rule engine. It is free software, released under the Apache License. The Drools engine is implemented in Java, as is JBoss, and is also controlled using that language. Initialization of the engine and deployment of rules is implemented in Java. Also, as unusual for production rule engines, rules never fire by themselves, but are issued to do so by the Java program that controls the engine. In addition, Drools can be extended by defining so-called Domain Specific Languages. These are languages that may have a different syntax than the standard Drools syntax to write queries in. Rules in Domain Specific Languages are then translated into the Drools language when inserted into the engine.

XChangeEQ is a research Event Query Language. It is developed at the University of Munich and designed for automated reasoning on the Semantic Web. XChangeEQ introduces a new style of event querying. It separates event query features into four so-called dimensions: Data extraction, event composition, temporal relationships, and event accumulation. Most operators belong to exactly one of these dimensions. This was done to define clear semantics. As XChangeEQ is designed for use on the Web, it works best at processing tree-structured events, such as XML messages. Queries are generally structured like the XML representations of the events queried. For querying simple events, it embeds the Xcerpt language. Xcerpt queries apply patterns to XML documents, similar to templates. Change is a reactive programming language. Using Event-

Condition-Action rules, it allows Web sites to react to changes at other Web sites, for example by updating its own data.

TelegraphCQ, from the University of California at Berkeley, is designed to provide event processing capabilities alongside relational database management capabilities by utilizing the PostgreSQL open source code base. The existing architecture of PostgreSQL is modified to allow for continuous queries over streaming data. Several components of the PostgreSQL engine underwent very little modification, while others were significantly changed. The most significant component of the TelegraphCQ system is the ”wrapper,” which allows for data to be pushed or pulled into the Telegraph processing engine, and custom wrappers allow for data to be obtained from any data source .

BEA Systems in 2007 introduced of their WebLogic Real Time and WebLogic Event Server systems. More specifically, their Event Server technology is a focus on event-driven service oriented architecture which provides a response to events in real-time. As part of the package, they provide a complete event processing and event-driven service-oriented architecture infrastructure that supports high-volume, real-time, complex, event-driven applications. This is one of the few commercial offerings of a complete, integrated solution for event processing and service-oriented architectures.

Truviso is a commercial event processing engine that is based on the research toward the Telegraph CQ project at UC Berkeley. The”claim to fame” for Truviso is that it supports a fully functional SQL, and integrates PostgreSQL relational database alongside a stream processing engine. The integration of PostgreSQL leads to other aspects of the Truviso system. The queries are simply standard SQL with extensions that add functionality for time windows and event processing. Carried over from PostgreSQL are user-defined functions, as well as JDBC and ODBC interfaces. In addition, the use of an integrated relational database allows for easy caching, persistence, and archival of data streams, as well as queries that include not only real-time data, but also the historical data [9].

3 Methodology

The development of the system followed the SDLC development methodology.In addition, object oriented approach was used in design and implementation.

3.1 Requirement Gathering

3.1.1 Interviews

Interviews were conducted with investigative officers at Police headquarters, Jinja Road police station and the Special Investigation Unit, Kireka. This was carried out to ascertain the nature of financial crime cases, the investigation processes involved and the criteria used to zero down on a suspect. Telephone interviews were also conducted in situations where physical access was challenging.

The questions asked were related to the following:-

1. Type of evidence gathered and used in such investigations.
2. The patterns which are common in most cases and which the system will base upon.
3. Any existing systems used for crime investigation and formats of existing crime records.

3.1.2 Existing Literature

Existing literature on Complex Event Processing was read to determine the best design approach for the project and to discover the benefits of using the CEP model.Implementations of the Esper engine in particular were studied to provide a benchmark for the project implementation.

The Annual Police report was also read to acquire an in depth understanding of the impact of crimes to society and government. Case statements some handwritten were also analyzed and these provided guidance on data to capture and track in the system.

3.1.3 Questionnaires

Questionnaires were distributed to investigation officers and their superiors to determine the need for such a tool and the benefits it would bring to the Police force. The returned forms portrayed a clear need for the system as it would reduce on processing time and reduce case backlog.

3.2 Requirements Review

Data gathered through the above data collection methods was sorted to enable selection of material relevant to the design process of the system. This involved determining the sources of input to the system, the format of incoming data events, the crime patterns to be checked against and the output format of the matching events in form of reports.

4 The Crime Investigation Tool

The system consists of four components at the highest level. These are the web client, processing engine (Esper), data source (MySQL Database) and application server (glassfish).

4.1 System Design

The design includes a web user interface which will enable crime investigators to select records of particular cases to be sent to the engine for processing and also to receive processed responses on web pages. The client communicates with the database to generate input data and interacts with managed beans (business logic) to send and receive events from the processing engine.

The records from the data source are sent to a processing engine that filters incoming events based on defined crime patterns. Records generated from the database and converted to events provide input to the engine. The engine sends regular responses to the web client showing events that have matched the described pattern. The engine implements a listener which constantly receives events as they stream in. This is works closely with application server that manages the deployment, compilation and running of the application. The crime pattern is determined by the investigator and can be adjusted through the user interface.

Records requested by the client are generated in CSV (Comma Separated Values) format and sent to the engine for processing. The records are stored in a MySQL Database for all cases that are being investigated. The events are represented in POJO (Plain Old Java Object) format with getters and setters in a form understandable to the CEP engine.

4.1.1 Data flow diagram

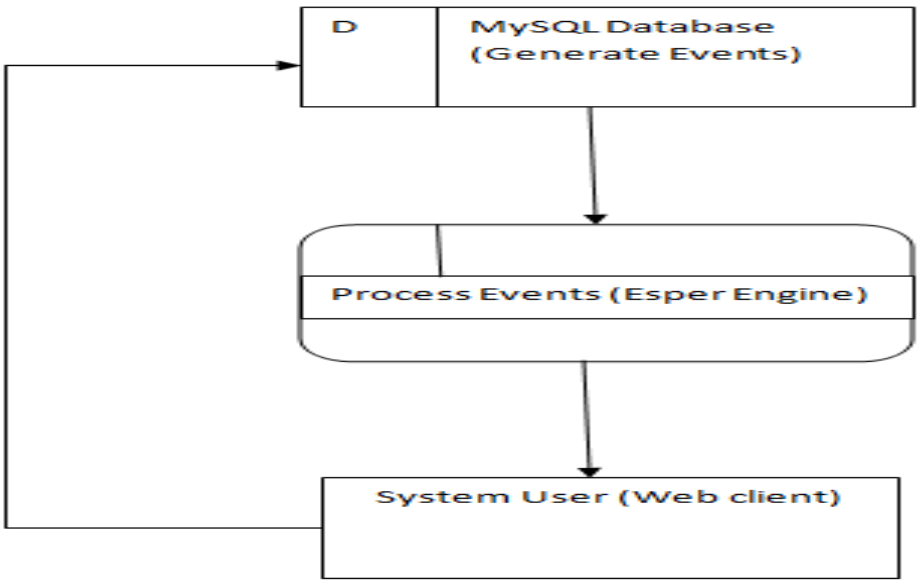

Figure 1: Data flow Diagram

4.2 System Implementation

The tool is a JavaEE6 web application was developed using the Netbeans 7.2.1 IDE (Integrated Development Environment) with the JSF 2.1 framework. A presentation-oriented web application generates interactive web pages containing various types of markup language (HTML, XHTML, XML, and so on) and dynamic content in response to requests.

The data source is a MySQL database that generates events on request from the user via the managed bean. The CEP engine used is Esper 4.8.0 with support CSV files and logging capabilities. JavaEE6 is a collection of APIs that simplifies the application development process and produces fast, reliable and secure applications.

The the client tier consists of dynamic web pages (XHTML) with capabilities to communicate with managed bean classes designed using the JSF 2.1 framework. Java Server Faces technology builds on Servlets and JSP technology and provides a user interface component framework for web applications.

The business tier/JavaEE Server contains JSF pages (servlets) and managed beans (POJOs) that describe the business logic and both reside on the JavaEE server. The Expression Language is

used to link the client web pages programmatically to the managed beans that expose getters and setters. This is where the Esper engine is configured and programmed to receive events and respond to the client with results.

Esper is an Event Stream Processing (ESP) and event correlation engine (CEP, Complex Event Processing). Targeted to real-time Event Driven Architectures (EDA), Esper is capable of triggering custom actions written as Plain Old Java Objects (POJO) when event conditions occur among event streams. It is designed for high-volume event correlation where millions of events coming in would make it impossible to store them all to later query them using classical database architecture.

A database was created using MySQL 5.5 to record and store all activities related to a case under investigation. This data is then retrieved from the database using the JPA API and converted to events by the managed bean before it is sent to the Esper CEP engine for processing. The JPA provides Java developers with an object/relational mapping facility for managing relational data in Java applications.

4.2.1 System 3-Tier Architecture

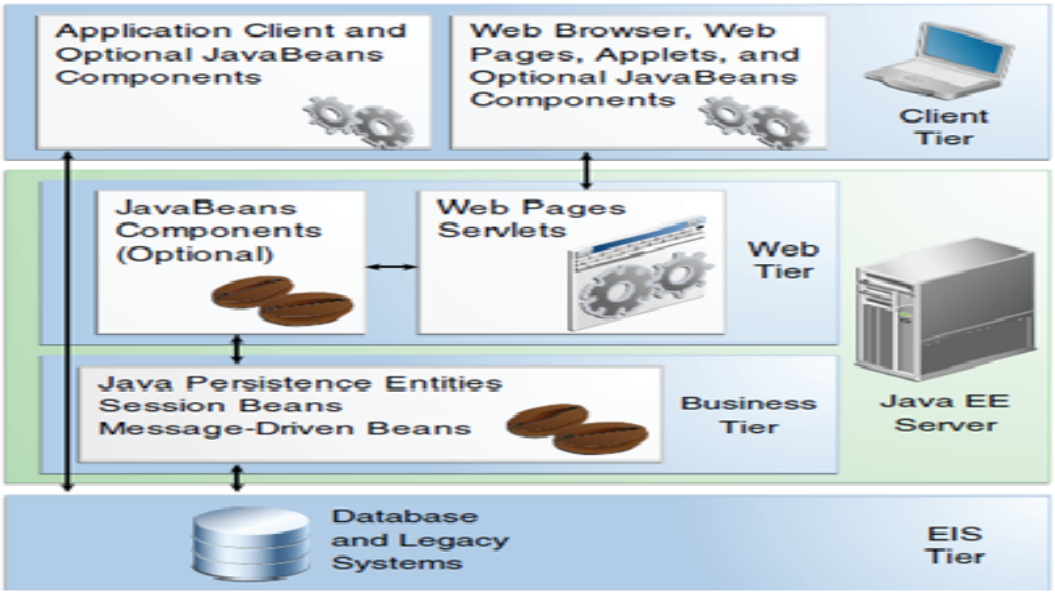

Figure 2: 3-Tier Architecture

4.3 Crime Patterns Considered

There are several patterns of behavior depicted by criminals involved in crimes and these differ according to the nature of crime.For purposes of this project, crimes related to misuse of funds are considered along with patterns realted to such crimes.

4.3.1 Patterns Related to Bank Account Activity

Most criminals involved in financially related crimes tend to possess large sums of money in their bank accounts that exceed their source of income. These amounts are deposited on a regular basis (daily, weekly, monthly)in different bank branches and are also used to acquire property within a short period of time.In investigating such crimes, investigators analyze company data detailing the earnings of an individual and the payment dates of his/her salary.They also probe assets acquired by the individual from the past todate detailing accquisition periods and amounts.

Other details investigated include:-

1. Bank Statements for the both the individual and company for selected periods.
2. Company budgets and financial statements.
3. External sources of income of the individual.
4. Financial authorization powers and signatory levels.
5. Funds transfered from accounts within and out of the country.

Theerefore the processing engine was configured to store patterns checking stremed events (bank statements, funds transfers,expenditures) against salary, known assets and external sources of income.