Data Streams, Complex Events, and BI

Seth Grimes

Alta Plana Corporation

+1 301-270-0795 -- http://altaplana.com

International Data Warehouse & Business
Intelligence Summit 2008
June 11-13, 2008

Alta Plana

Agenda

BI at the speed of thought → Analytics at the speed of data: the business case.

Technologies and Solutions for real-time data and real-time analytics.

Evaluating and implementing: Best Practices, strategy, and resources.

Alta Plana

Data Streams

I worked for 3 years for NASA...

Programming simulations of satellites that would map the earth's gravity field.

Data involved simulated orbit altitude, inclination, duration, and the accuracy and measurement rate of instruments.

http://media.skyandtelescope.com

Instruments captured acceleration with 6 degrees of freedom – 3 spatial dimensions plus "attitude": roll, pitch & yaw.

I also worked on magnetic-field mapping where satellites would be ranged by ground-based lasers.

Data Streams

Then I worked for 5 years for the U.S. National Highway Traffic Safety Administration...

Vehicle crash tests (at that time) included up to 80-90 sensors on vehicles and dummies, 8000x/second data rate.

Data was correlated to video.

http://www.nhtsa.dot.gov/

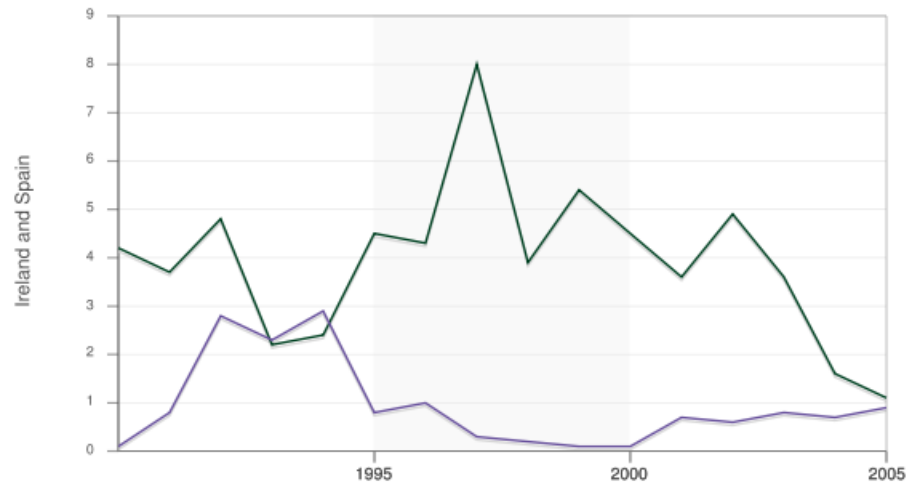Analyses were off-line.

Alta Plana

Data Streams

I then worked with economic data, especially time series...

Observations are at calendar frequencies: yearly, monthly, weekly, etc.

Econometric models link disparate data and enable forecasting.

Alta Plana

Data Streams

Data streams come from many sources.

Data volumes, types, frequencies vary.

Processing approach depends on use:Data-stream correlation; analysis techniques.

Communications networks and capital markets created the market for new classes of software.

Data stream processing (DSP).

Complex event processing (CEP).

Builds on BI and Operations Research.

OR=mathematical modelling for automated decision making.

Data Streams

Capital markets involve trades, bids/offers, news, statistics (e.g., P/E ratio, float), competitive intelligence, and economic & market data.

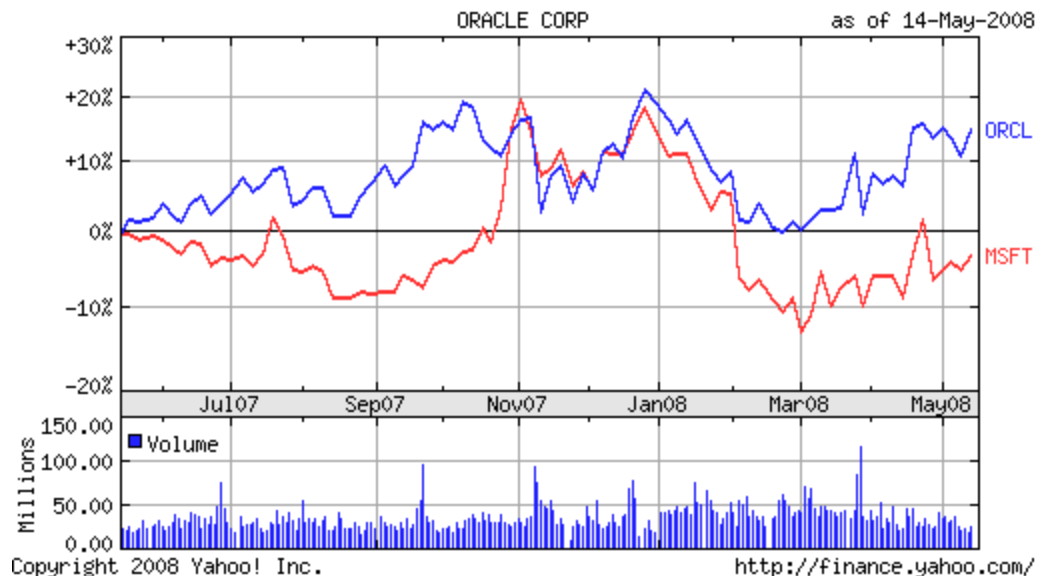Trading data includes symbol, volume, price, time.

Many derivatives:

Put & call options.

Indexes & baskets.

ForEx hedging.

Monetizations.

Breakthrough Analysis

Traditional BI is not enough.

Historically focused rather than "operational" and predictive.

Designed for reporting and interactive analysis rather than automated, embedded analytics.

Issues with real-time (RT) data and real-time response.
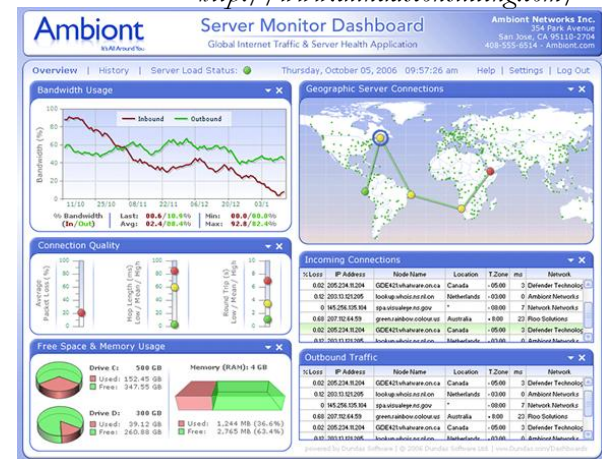
Tilt toward performance management, financials.

http://www.dundasconsulting.com/

On the plus side, BI has:

Dashboards.

KPIs.

Ability to handle time-based data.

Growing RT capabilities.

International BI-DW Summit

Breakthrough Analysis

Traditional DBMSes and DW approaches –

   Load → Index → Query

– do not meet the speed need.

Standard SQL and standard SQL engines don't support continuous queries.

Common APIs (interfaces) – ODBC/JDBC, Analysis Services – are set/record oriented and do not provide adequate real-time query capabilities.

Breakthrough Analysis

We need to cope with –

> Bigger data.

> Faster data.

> New, complex information types.

> Low-latency ("real-time") processing.

> Analysis over data streams.

> Complex, statistically rooted algorithms.

> Automated decision making.

=> Complex Event Processing.

Problem Statement

More data sources, more & faster data:

- RFID (radio frequency identifier) chips, especially for inventory and logistics.

- News and social media.

- Retail transactions.

- Communications networks.

- Sensor networks, security devices.

- Trading markets.

Analytics driver: efficiency, speed, accuracy.

Problem Statement

More data types:

Surveillance video and audio.

Sensor state information (e.g., temperature, humidity).

Geolocation data, e.g., from GPS devices.

Complex transactional information.

Bids and offers.

Full ("unstructured") text.

Coded and tagged ("structured") text.

Timestamps.

Problem Statement

Challenges:

Timely response.

Capture of high-frequency data.

Correlation of data from multiple sources.

Long-running and time-aware queries.

Issues:

RDBMSes were designed for transaction processing.

Load, structuring, and indexing time given traditional datamarts and data warehouses.

Analytical styles: Traditional BI is historically focused.

Algorithms: Scalability of traditional serial algorithms.

Responses

Operational/Real-Time BI?

Data Stream Processing (DSP).

Complex Event Processing (CEP).

...

Adaptive Data Warehousing.

Column stores.

Massively Parallel (database) Processing (MPP).

In-memory analytics.

Operational/Real-time Data/BI

Real Time BI: Fast response required.

> E.g., model scoring.

> Supports automated decision making.

Operational BI on real-time data – *What's happening* **right now**? – involves a bit more.

> Dashboards provide a great visualization tool.

> BI differs from monitoring: it contextualizes information.

> Be careful of "twinkling" data.

> But cleansing, structuring, indexing add latency.

> And **we're still not handling events or complexity**.

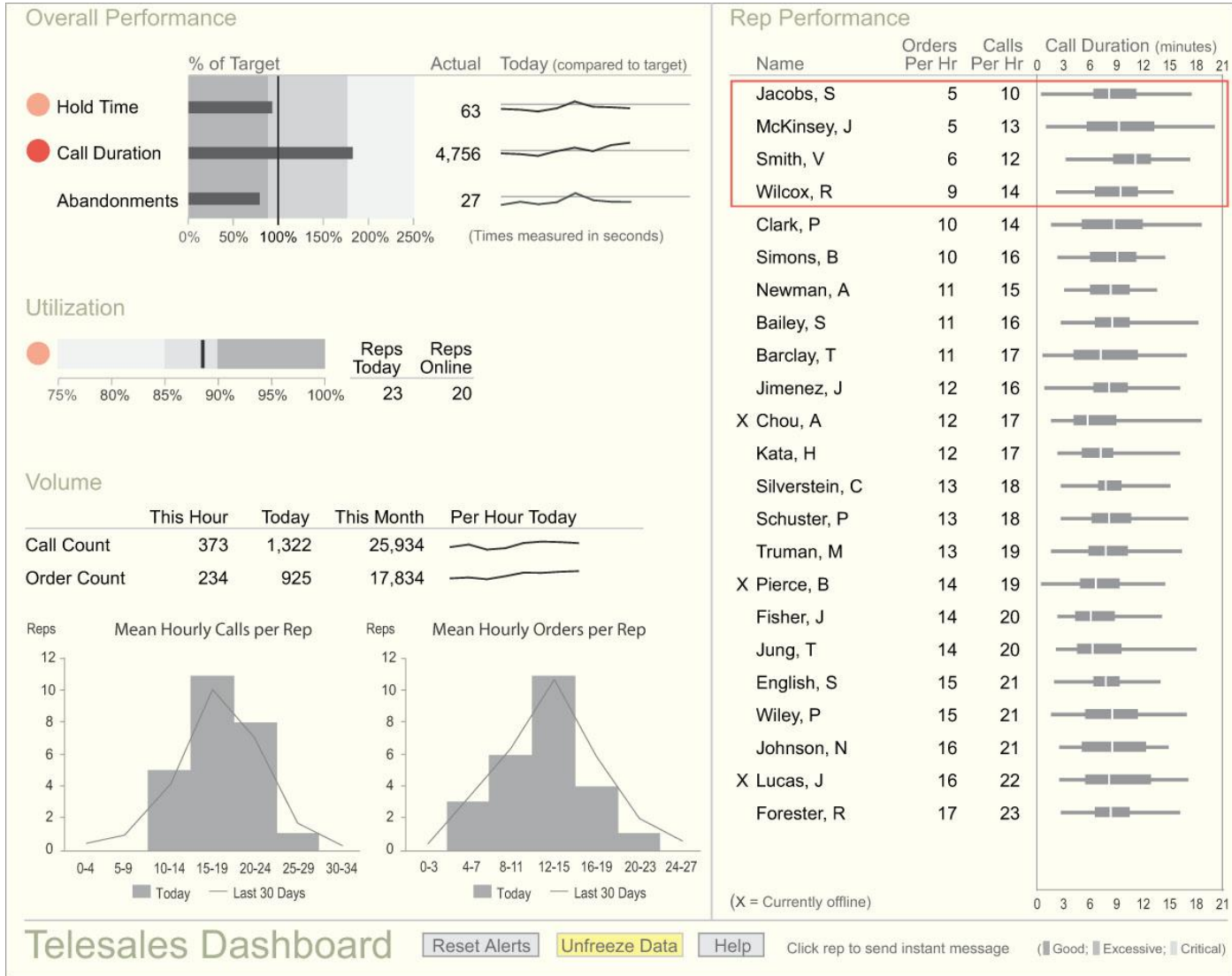# Operational/Real-time Data/BI

Most BI dashboards are not for RT data.

http://www.perceptualedge.com/blog/?p=154

Events

The concepts of events became familiar in the '90s in association with object-oriented (OO) programming.

C++, Smalltalk, Java, and other languages.

Software "objects" are invoked in response to program events such as the user's pressing a button.

Objects and events are managed in an application server.

This conception did not cover **data** events.

We did have DBMS based event handlers: stored procedures, triggers, alerts, etc.

Alta Plana

Events

Data events:

- A shop in Bangkok seeks to authorize a charge on a credit card that was used one hour ago in Nairobi.
  - Inventory of widget X fall below five days' average sales.
  - The spread between the AU.L price in London and the AUY.F price in Frankfort exceeds 3%.

International BI-DW Summit
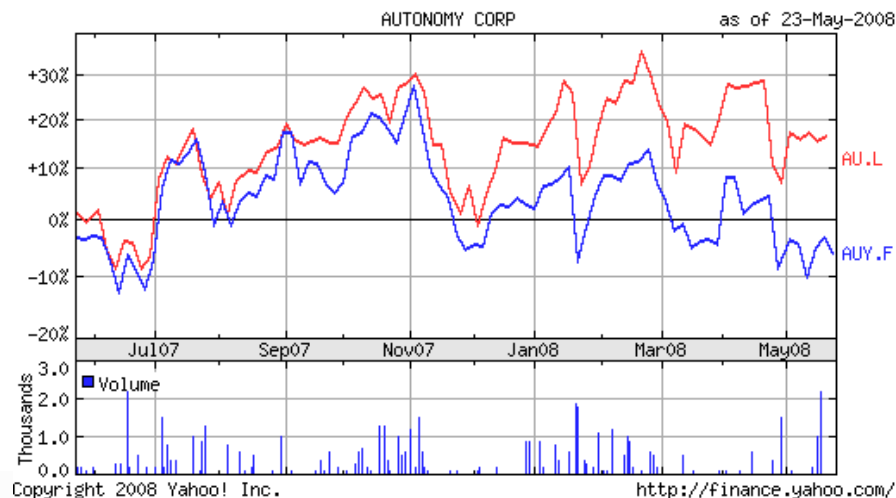
Data Stream Concepts

Ordering: A stream is a sequence of observations over time.

Correlation: Multiple streams may be related.

Time and observation windows.

Continuous/long-running queries.

Use of SQL, treating streams like tables.

Use of SQL joining streams and historical data.

Data Stream Concepts

"Traditional ETL processes are limited to reading data at rest: from databases, mainframes, and files extracted from other operational systems.

"*Data in flight* exists in other formats: messages on message-oriented middleware, web service calls, TCP network packets, and so forth...

"A stream is analogous to a table in a relational database; but … contains an infinite sequence of rows that arrive whenever the producer decides to send them."

    -- Julian Hyde, SQLstream architect
       (*http://julianhyde.blogspot.com/2008/02/streaming-sql-meets-olap.html*)

Data Streams and CEP

Extend the concept to event streams.

My definition: An "event" involves a data condition that is notable in some context.

Those BI dashboards showed events (and more) but don't handle streams.
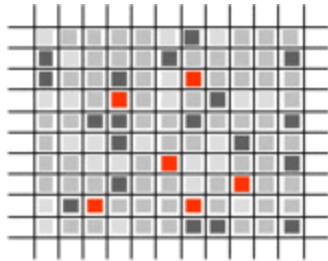
Complex Event Processing.

Feeds on data and event streams and "clouds."

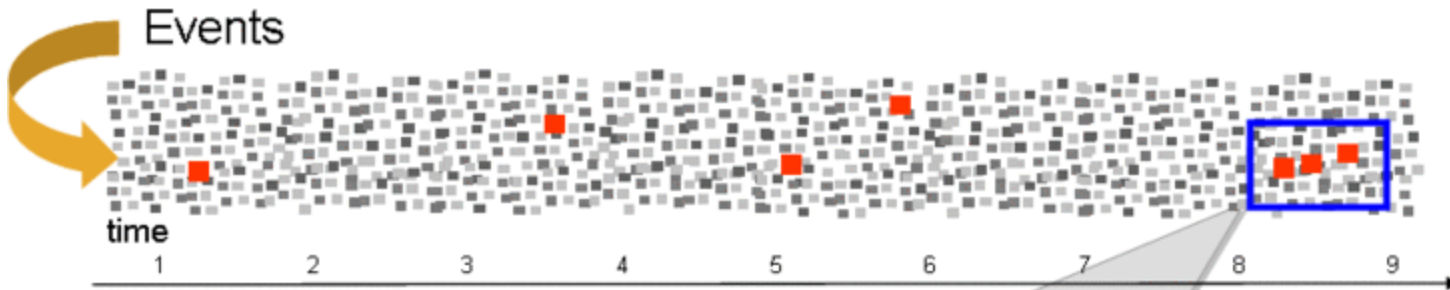Transforms input streams – whether data are atomic or complex – into output streams.

Supports automated decisions via rules systems.

BI versus CEP

Copyright 2007 Progress Software

Static Data Processing: "What potentially abusive trades happened last week?"

Events

time

1   2   3   4   5   6   7   8   9

Event Processing: "When a trader buys and sells the same security either simultaneously or within a relatively short period of time, alter fraud for wash trading"

Alert

http://www.eventstreamprocessing.com/

Data Stream Pioneers

AT&T Labs created Hancock: A Language for Extracting Signatures from Data Streams, in 1999-2000.

"Massive transaction streams present a number of opportunities for data mining techniques. Transactions might represent calls on a telephone network, commercial credit card purchases, stock market trades, or HTTP requests to a web server. While historically such data have been collected for billing or security purposes, they are now being used to discover how clients use the underlying services."

http://www.research.att.com/~kfisher/hancock/kdd.pdf

Data Stream Pioneers

Stanford University:

2001: Continuous Queries over Data Streams.

STREAM: The Stanford Stream Data Manager, 2002-6.

CQL Continuous Query Language.

Brown/Brandeis/MIT:

Aurora Project: "The fact that a software system must process and react to continual inputs from many sources (e.g., sensors) rather than from human operators requires one to rethink the fundamental architecture of a DBMS for this application area."

(http://cs-www.cs.yale.edu/homes/dna/papers/vldb095.pdf)

Borealis Distributed Stream Processing Engine follow on.

Data Stream Commercialization

Coral8: Rajeev Motwani from STREAM.

StreamBase: Jennifer Widom from Stanford plus Stan Zdonik and Michael Stonebraker from Aurora/Borealis.

Stonebraker's entry validated the field. He:

- Created the Ingres RDBMS.
- Created the Postgres ORDBMS.
- Commercialized Postgres as Illustra, sold to Informix.
- Co-founded StreamBase.
- Created Vertica, a leading column-store analytical DBMS.
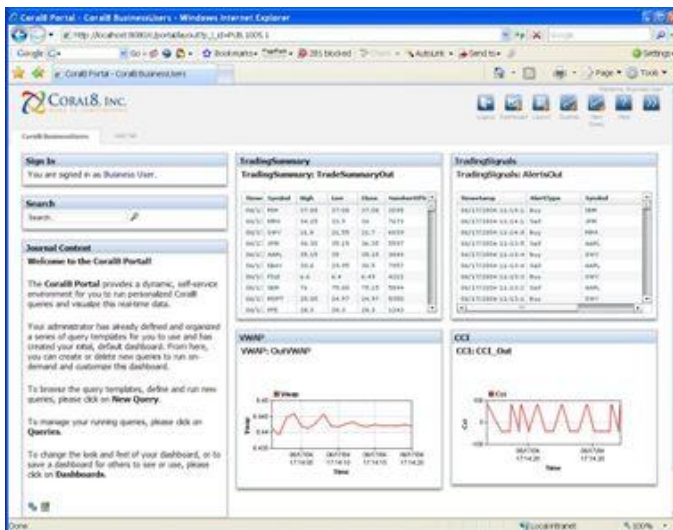
CEP Ingredients

Take Coral8 as an example.

Developer focused, providing application building blocks and an Eclipse IDE plug-in.

Continuous Computation Language (CCL) extends SQL with windows, time series operations, and pattern matching.

Accepts XML formatted data.

Adapters for common data feeds such as Reuters.

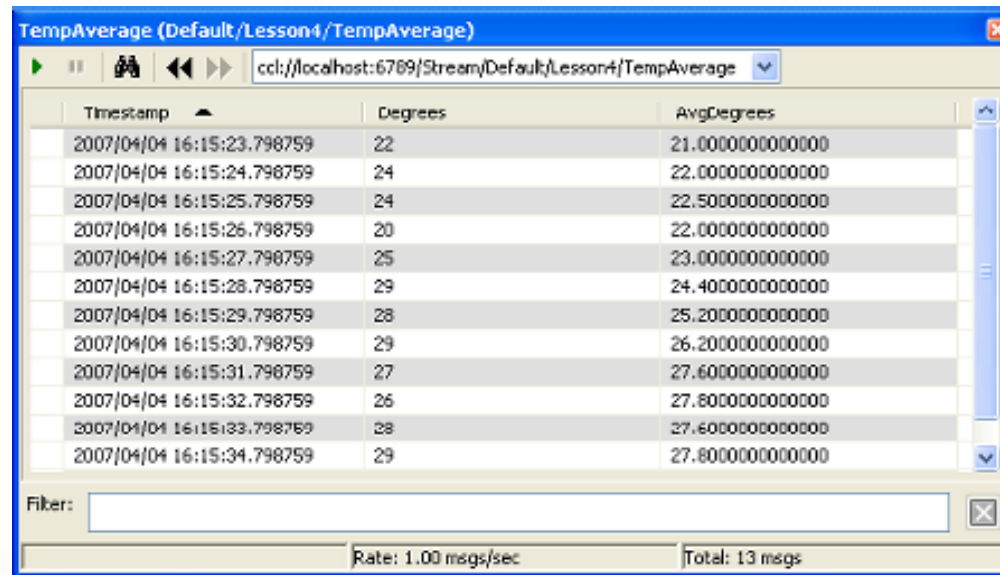Portal interface for RT monitoring and visualization; Microsoft Excel integration.

http://www.coral8.com/products/portal.html

CEP Ingredients

CCL examples: *(http://www.coral8.com/system/files/assets/pdf/GettingStartedWithCoral8.pdf)*

Create an output stream, TempAverage, from a 5-row sliding window extracted from an input stream, TempOut:

```
INSERT INTO TempAverage
SELECT Degrees, AVG(Degrees)
FROM TempOut KEEP 5 ROWS;
```
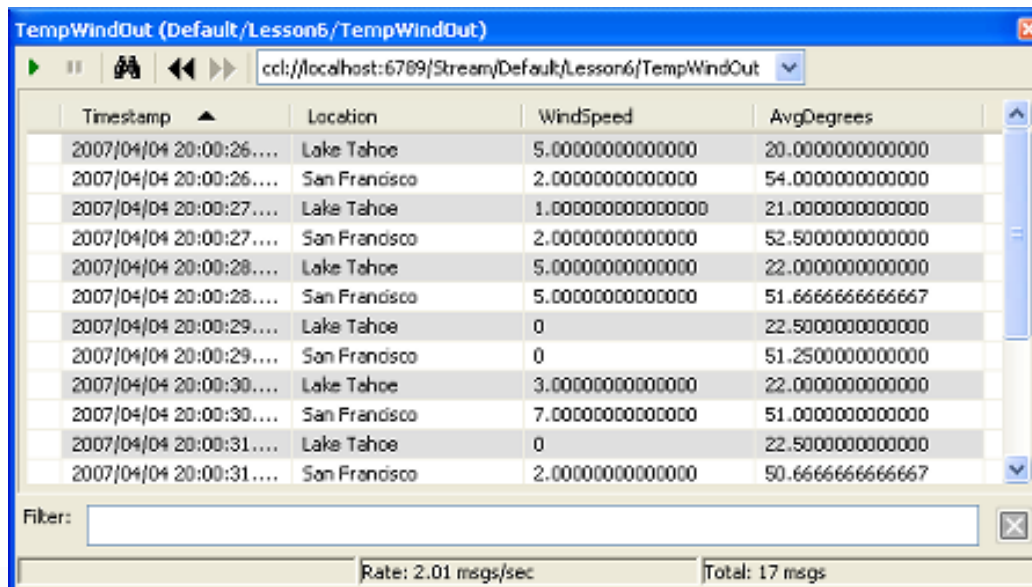

TempAverage (Default/Lesson4/TempAverage)

ccl://localhost:6789/Stream/Default/Lesson4/TempAverage

Timestamp	Degrees	AvgDegrees
2007/04/04 16:15:23.798759	22	21.000000000000
2007/04/04 16:15:24.798759	24	22.000000000000
2007/04/04 16:15:25.798759	24	22.500000000000
2007/04/04 16:15:26.798759	20	22.000000000000
2007/04/04 16:15:27.798759	25	23.000000000000
2007/04/04 16:15:28.798759	29	24.400000000000
2007/04/04 16:15:29.798759	28	25.200000000000
2007/04/04 16:15:30.798759	29	26.200000000000
2007/04/04 16:15:31.798759	27	27.600000000000
2007/04/04 16:15:32.798759	26	27.800000000000
2007/04/04 16:15:33.798759	28	27.600000000000
2007/04/04 16:15:34.798759	29	27.800000000000

Filter:

Rate: 1.00 msgs/sec   Total: 13 msgs

CEP Ingredients

Join one stream with 10-second sliding time window extracted from a second stream:

```
INSERT INTO TempWindOut
SELECT WindIn.Location, WindSpeed, AVG(Degrees)
FROM WindIn, TempIn KEEP 10 SECONDS
WHERE WindIn.Location = TempIn.Location
GROUP BY WindIn.Location;
```

CEP Ingredients

Use an event:

```
INSERT INTO WindPatternOut (Location,Speed1,Speed2)
SELECT W1.Location, W1.WindSpeed, W2.WindSpeed
FROM WindIn W1, WindIn W2
MATCHING [2 SECONDS: W1 && W2]
ON W1.Location = W2.Location
WHERE (W1.WindSpeed - W2.WindSpeed) >= 5;
```

CEP Ingredients

Tools typically offer modelling interfaces.

http://www.aleri.com/products/aleri-streaming-platform/high-level-authoring-tools/
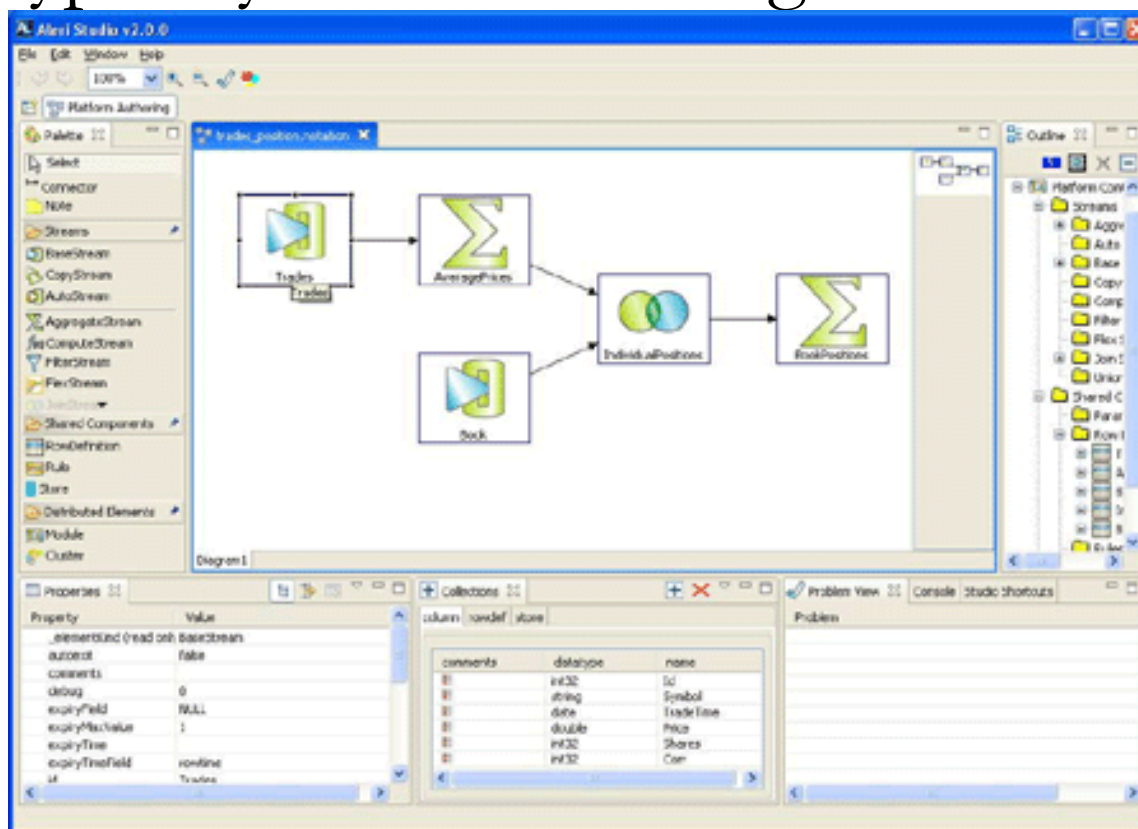
CEP & Event Driven Architectures

We've seen a CCL construct for defining events. How does CEP fit with Event Driven Architecture (EDA)? Some definitions:

Service-Oriented Architecture (SOA): Business processes are defined, published, and invoked as services.

SOAP is an example of a Web-services protocol, typically involving remote procedures calls (RPCs) over HTTP using XML.

Enterprise Service Bus (ESB): A mechanism, typically a framework involving messaging infrastructure, for inter-connecting services in an SOA

CEP & Event Driven Architectures

CEP builds events from streams and "clouds."

CEP uses Event Processing Agents to take action according to rules.

- Whenever three timeouts have happened send an alert to the network manager.
- If more than ten shopping carts have been active for more than five minutes then activate the website reaction time monitor and display an amber alert on the dashboard.
- Whenever IBM trades 2% above its 1 hour VWAP (Volume-Weighted Average Price) and then within 15 minutes trades 5 points below then buy 1000 shares IBM.

http://complexevents.com/?p=195

CEP & Event Driven Architectures

CEP solutions integrate input from diverse, "cloud" sources and formats such as –

Data streams.

RDBMSes via queries and APIs.

Message queues.

RSS feeds.

XML.

Structured data files, e.g., comma-separated value (CSV).

Web clickstream and network logs.

E-mail and instant-messaging traffic.

Business Event Clouds
Adapted from Roy Schulte, Gartner

Large Companies Experience 10^4 to 10^7

Business Events Per Second

Traditional OLTP
Airline reservations
ATM transactions
Insurance claims
Orders
Inquiries

Sensor Networks
Device & application alerts
Bar code scans
Factory floor
RFID networks

Other Transaction Streams
Telco Call Records
Financial trade "ticks"
Credit card authorizations and transactions
Micropayments

Web and Net
Web commerce
Web 2.0 & social nets
Click streams
Poll Web page scrapes
RSS, Atom feeds
e-mail, IM

www.thecepblog.com

Case Study: Capital Markets

A focused, well-defined, high-value problem...

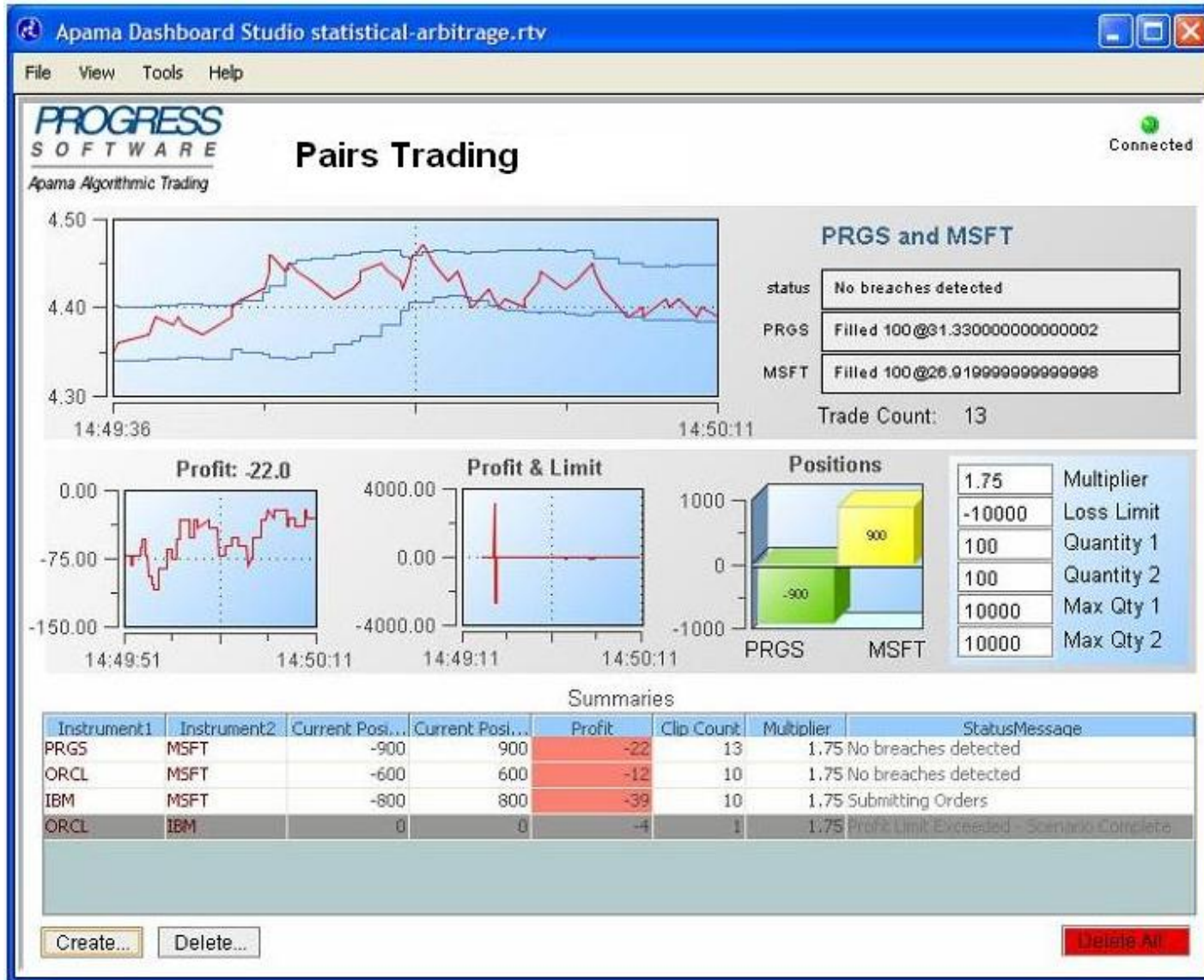Opportunity: Pricing anomalies derived from unequal information.

Solution: Arbitrage, algorithmic trading.

Problem: Latency (delay) – data & computations.

Challenge: Correctness, speed. Rapid execution=competitive edge.

Solution: CEP, automated trade execution.

Alta Plana

International BI-DW Summit

Case Study: BAM

Business Activity Monitoring.

Taps into operational systems, typically via SOA.

Real-time analysis and presentation of real-time data.

Computation of performance indicators.

Dashboard presentation.

CEP adds (Tim Bass, *cepblog.com*) —

Process monitoring.

Event tracking.

Situation detection

Predictive analysis.

Business Activity Monitoring

http://www.eventstreamprocessing.com/
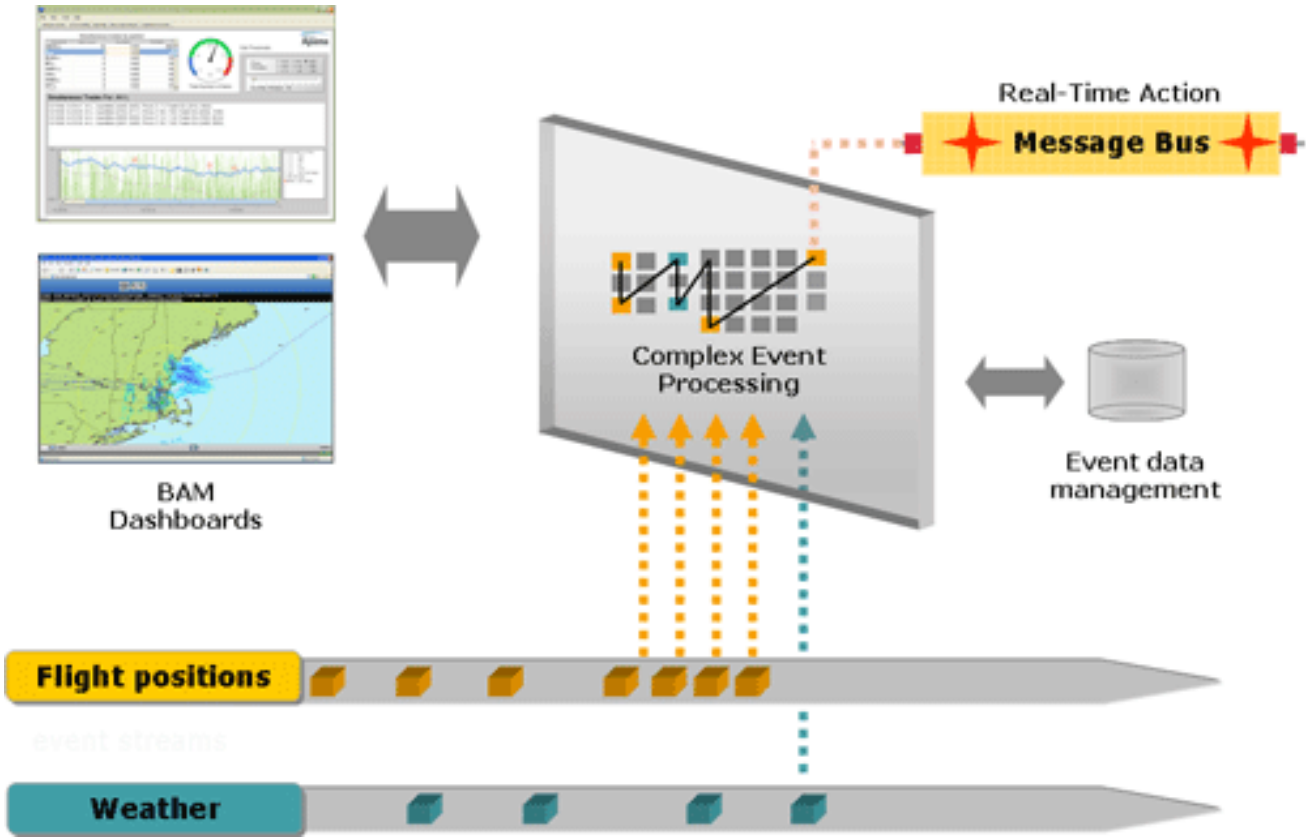
CEP Providers

Some CEP providers:

Aleri

Coral8

IBM/Aptsoft

Progress/Apama

Skyler

SQLstream

StreamBase

TIBCO

Truviso

Vhayu

Open Source CEP: Esper

Esper –

    Supports joins to historical data in a variety of RDBMSes.

    GNU General Public License v2; commercial license available.

    Available for Java as Esper and for .NET as NEsper.

    EsperHA (high availability) is embeddable.

    Does not support Java Message Service (JMS).

http://esper.codehaus.org/

CEP and the Data Warehouse

Anant Jhingran is CTO of IBM Information Management *(http://www.coral8.com/products/DB2.html)* –

"The close–knit relationship between Complex Event Processing engines and database management systems requires tight integration between the respective development languages and run-time servers."
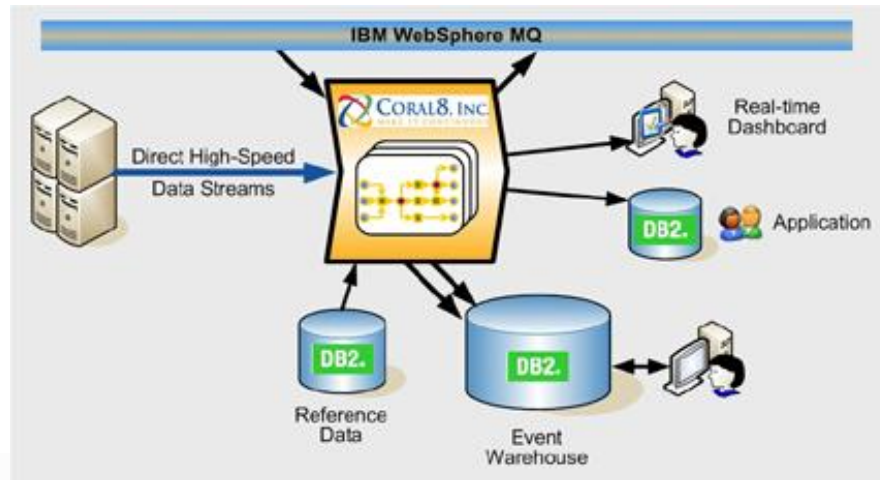
CEP and the Data Warehouse

Other partnerings –

Coral8 – Kx Systems kdb+ database for tick data.

StreamBase with Vhayu, again for low-latency access to tick data.

StreamBase with Vertica, the column-store DBMS.

CEP Extending the DBMS: Truviso

Truviso extends the PostgreSQL open source database management system.

"The new feature enhances PostgreSQL's performance for extremely long-running queries and transactions"

Single-server benchmark of 200,000 records/second.

PostgreSQL BSD license lets you use, modify, and distribute in any form you like, open or closed source.
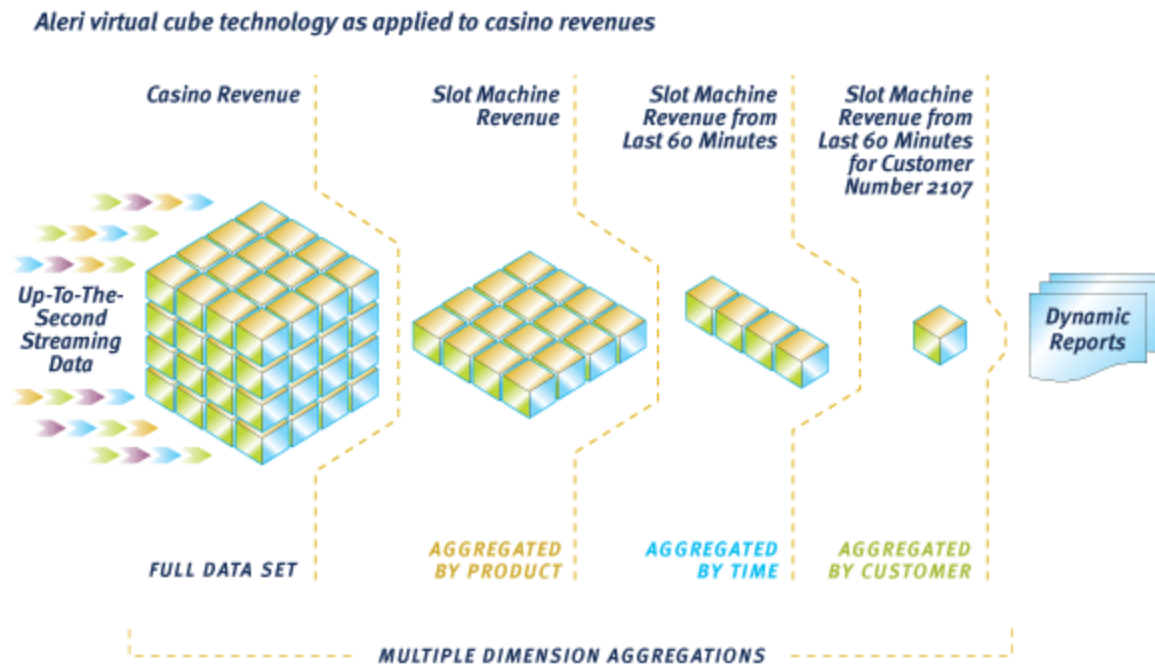
Extensions slated for inclusion in PostgreSQL release in early 2009.

Truviso includes integration and UI-building frameworks.

http://truviso.com/news/080418.html

Alta Plana

International BI-DW Summit

CEP and BI-Style Analysis: Vhayu

Vhayu offers another possibility, Live OLAP:

Aleri virtual cube technology as applied to casino revenues

Complex Event Processing

Is CEP enough? Is it too much?

Lack of understanding of "streams" and "clouds" and the value of the data the contain.

Problem of integrating historical-data query.

Entrenched approaches & technologies, e.g., data warehouses, Oracle and SQL Server DBMSes.

Entrenched BI tools and work practices.

Insufficient perception of need. Users may be waiting for capabilities to surface in their existing tools.

Alta Plana

Complex Event Processing

Parallelized (MPP) and column-store DBMSes, in-memory technology, and rules engines can present an alternative...

For focused, non-"cloud" problems.

Otherwise, view them as complementary.

Avoid custom coding (Java, C++) in today's fast-moving, diverse computing world.

Alta Plana

Column Stores and MPP

Columnar databases:

No indexing.

Compression.

Less data touched in queries.

InfoBright, ParAccel, Sybase IQ, Vertica, others.

MPP:

"Shared-nothing," parallelized architecture.

"Horizontal partitioning" with distributed processing.

Also, usually, no indexes: table scans instead.

DATAllegro, Greenplum, Netezza, Teradata, column-stores.

 International BI-DW Summit

The Way Forward

Investigate use cases – possible applications – at your organization.

- Data streams and cloud sources you're not now using: devices, log file, databases, data feeds.

- Events you can define and act on.

- Points where you can automate decision making & execution.

Explore implementing CEP *alongside* existing BI/DW solutions.

- Anticipate new data sources and analytical needs.

Evaluate SOA as an *enabling* architecture.

Resources

The Complex Event Processing blog.

http://thecepblog.com/

Prof. David Luckham's CEP portal.

http://complexevents.com/

Opher Etzion.

http://epthinking.blogspot.com/

Mark Palmer's CEPinterest.

http://www.eventstreamprocessing.com/

Resources

David Luckham, *The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems*

Questions?

Discussion?

Thank you!