See the wood for the trees

Dr. Harald Schöning
Head of Research

Get There Faster.™

The world is becoming digital

socienty

government

economy

Digital Society

Digital Government

Digital Enterprise

Data is Getting Bigger

Rapid Growth of Global Data from 2009-2020

From 1 to 35 **ZETTABYTES**

$125*10^{12}$ facebook friendship links 2012

Global mobile data traffic will surpass[3]

10 EXABYTES in 2016

RFID Market to see some serious growth[2]

$125*10^{9}$

RFID tags in 2020

The number of mobile-connected devices will exceed the world's population in 2013

$7*10^{9}$

Every day in the Internet [4]

12 TERABYTES Twitter tweets

24 PETABYTES processed by Google

software AG

Big Data is Largely Unexplored

5%

The ava

that firm

new business models
new products
new services

ttern discovery

automation

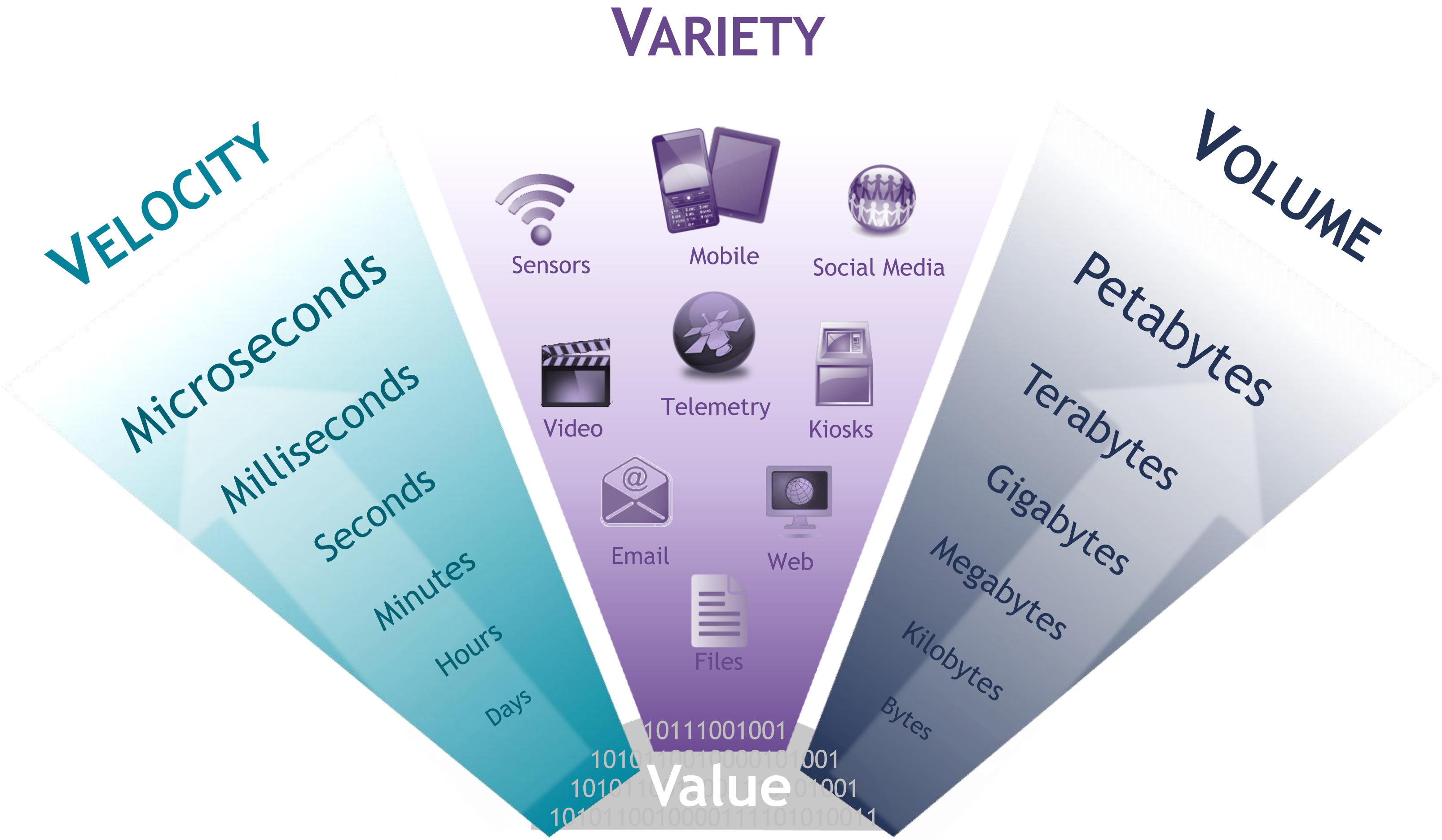
ictive analysis

Need discovery

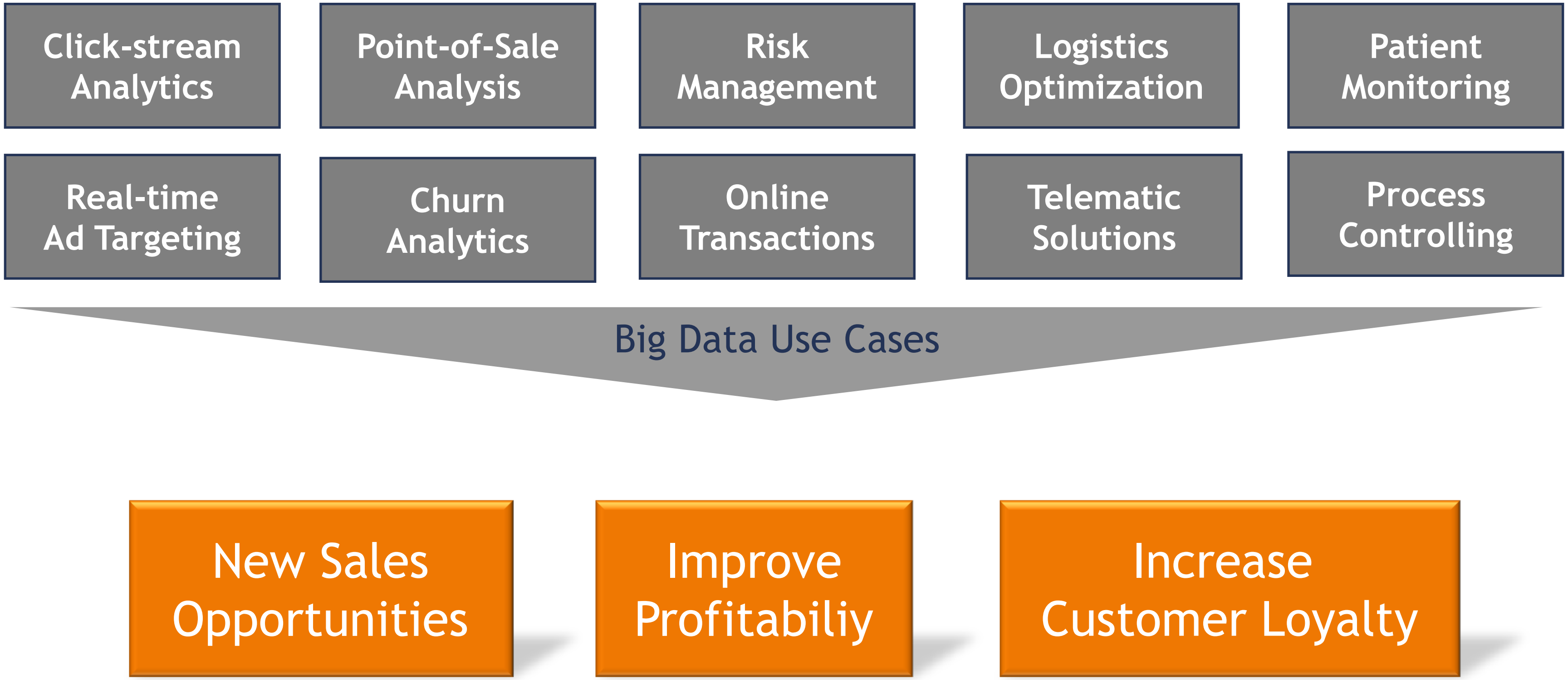
Automatic correlation

transparency

Big Data Phenomena



Big Data generates Business Value



Click-stream Analytics

Point-of-Sale Analysis

Risk Management

Logistics Optimization

Patient Monitoring

Real-time Ad Targeting

Churn Analytics

Online Transactions

Telematic Solutions

Process Controlling

Big Data Use Cases

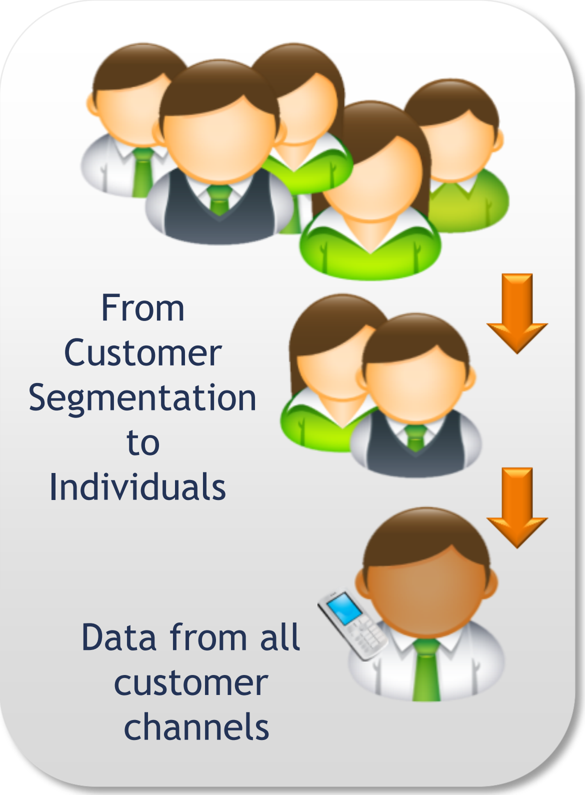
New Sales Opportunities

Improve Profitabiliy

Increase Customer Loyalty

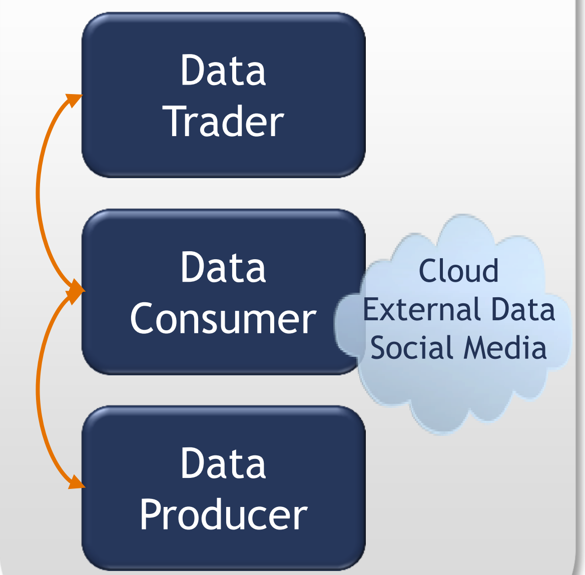
Data becomes more valuable

Customer Profiling


From Customer Segmentation to Individuals
Data from all customer channels

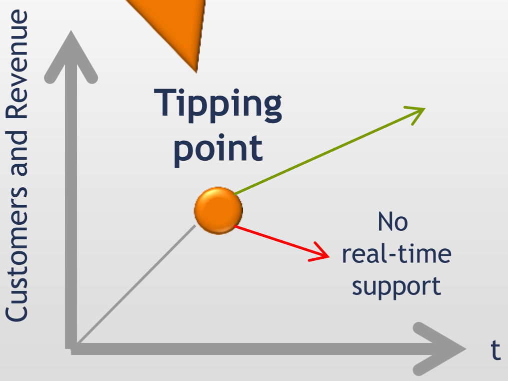
Data as trading good

From Data to Information



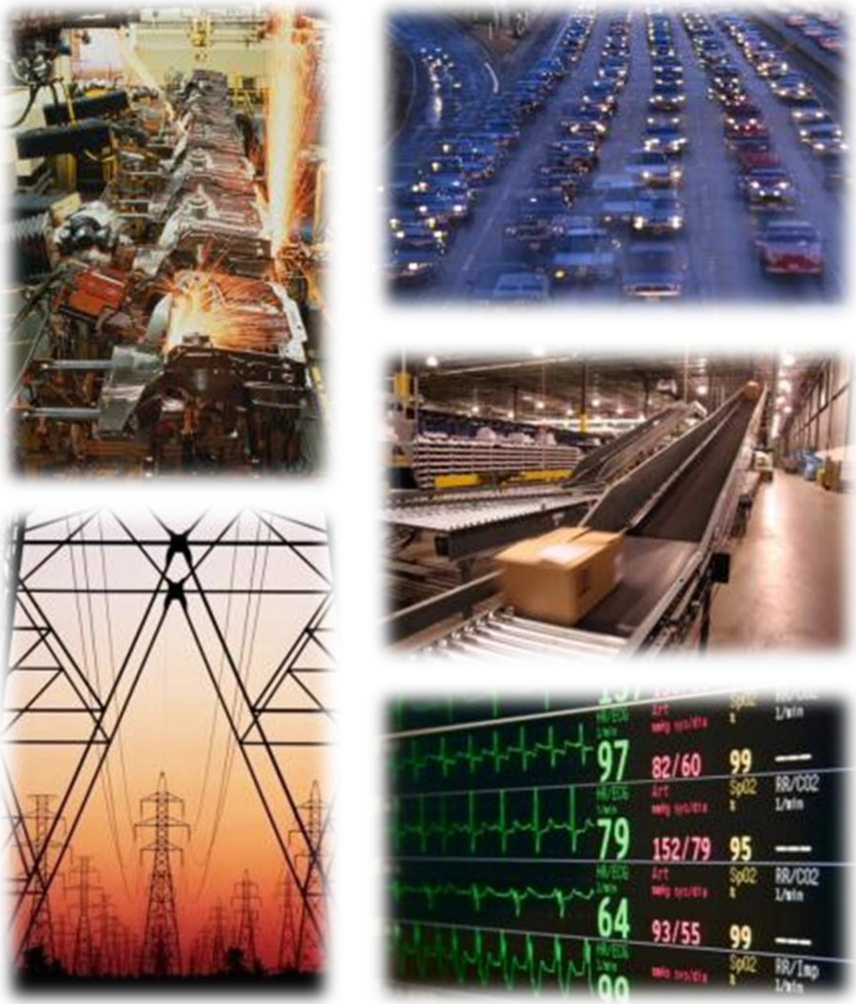
Real-time Business

When real-time is a game-changer in industries on loosing customers and revenue



Where do we face Big Data?

- Capital markets trading
- Fraud detection
- Logistics management
- Dynamic resource scheduling
- Service analytics & offers
- Incident management
- Smart metering & smart grids
- Governance, risk & compliance
- Supply chain automation
- Plant monitoring
- Traffic management
- Patient monitoring
- Transaction monitoring
- …



Use Cases - Overview

Manufacturing

- Track & Trace
- Transportation optimization
- RFID
- Internet of Things

Retail

- Customer Experience Management
- Real Time Coupons

Banking

- Fraud Detection
- Risk Mitigation
- Personalized Contextual Marketing

Energy / Utilities

- Demand Management
- Asset Management
- Outage Management
- Smart Grids
- Sensor Data Analysis

Life Sciences/ Healthcare

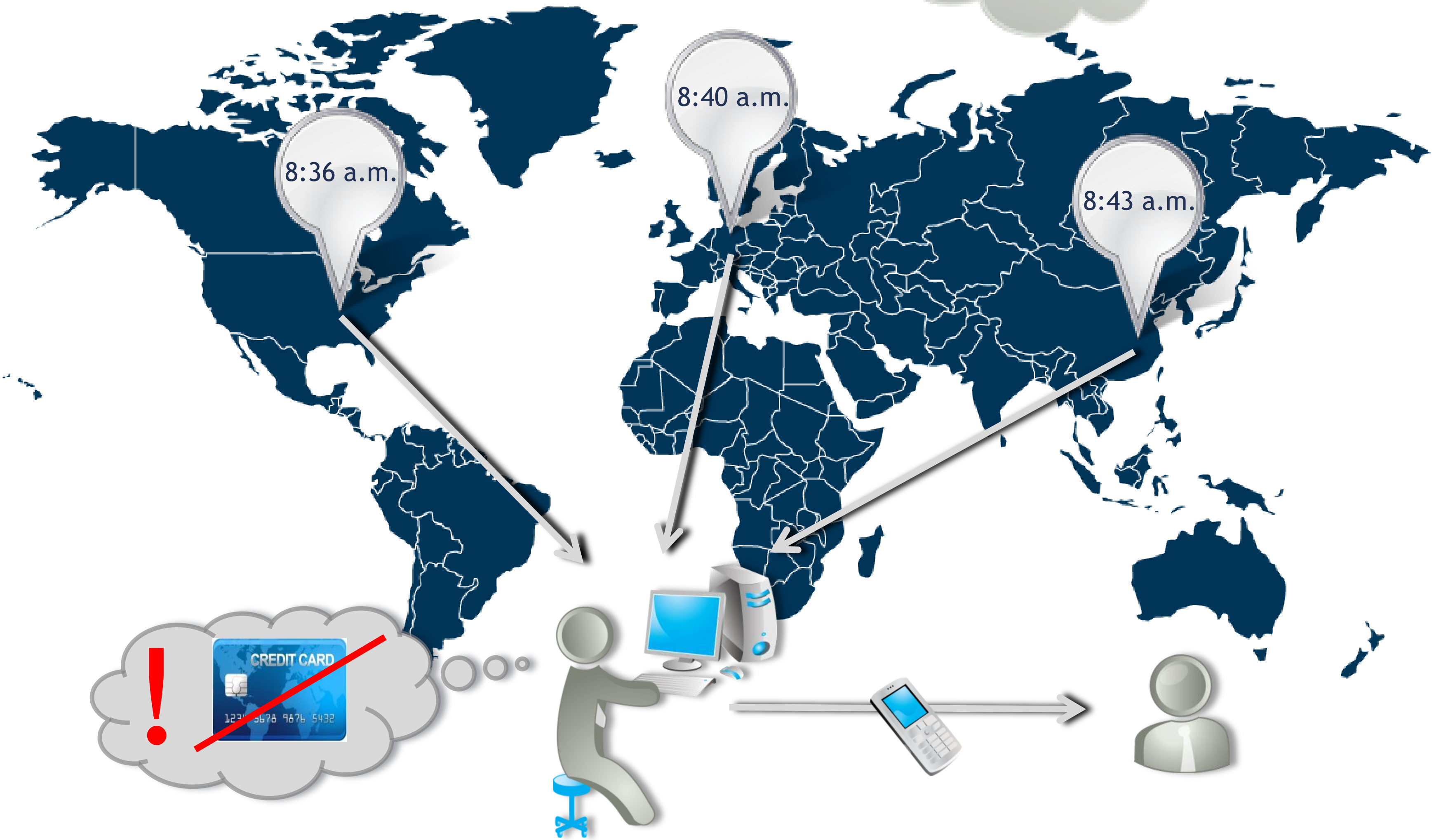
- Patient Logistics
- Home Health Monitoring
- Assisted Living

Government

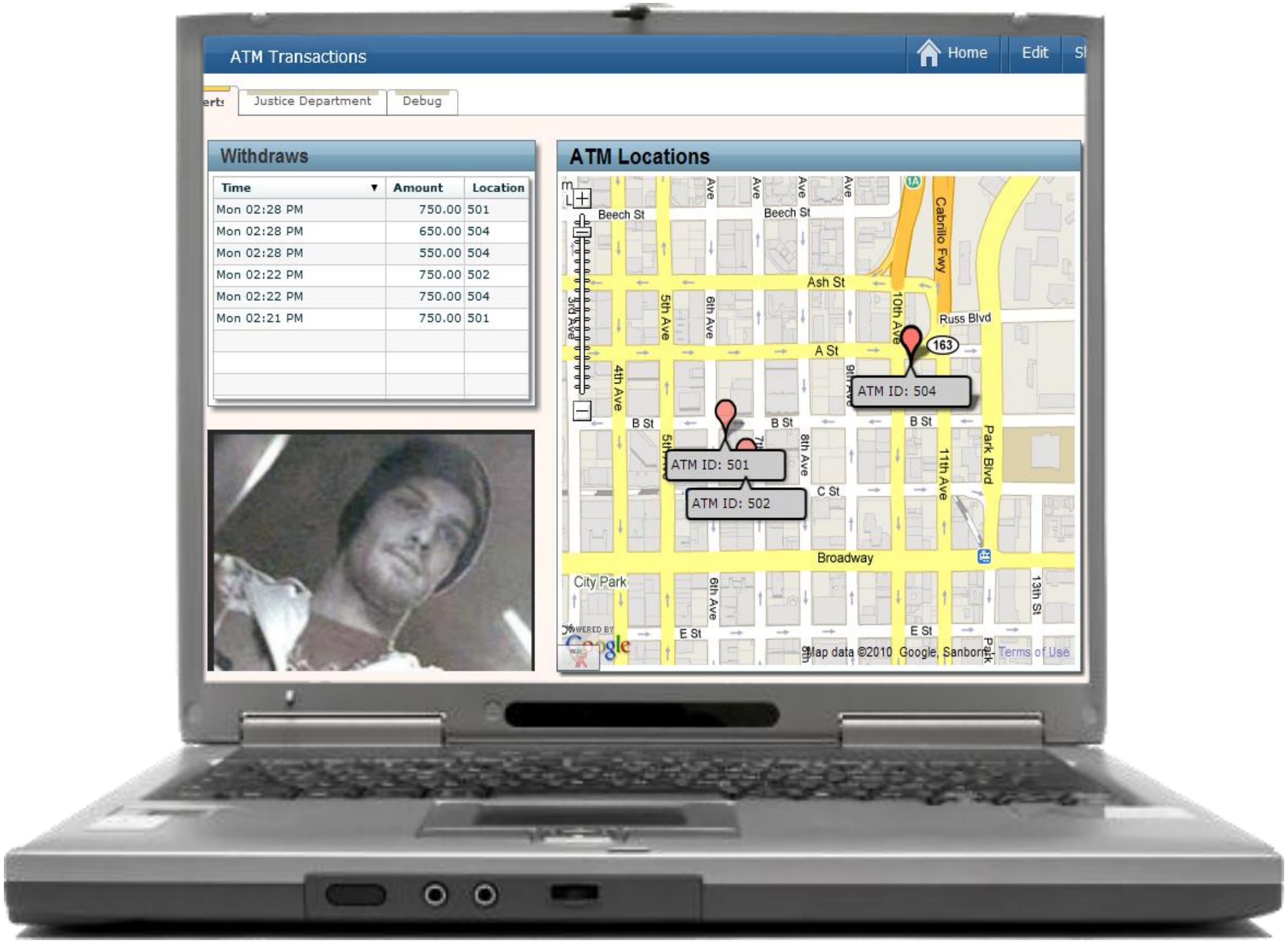
- Traffic Management
- Counter Terrorism

ATM Fraud Detection - Scenario

What would you do,
if you knew that…



Fraud Detection



Withdraws

Time	Amount	Location
Mon 02:28 PM	750.00	501
Mon 02:28 PM	650.00	504
Mon 02:28 PM	550.00	504
Mon 02:22 PM	750.00	502
Mon 02:22 PM	750.00	504
Mon 02:21 PM	750.00	501

ATM Locations

The Benefit of Preventing One Fraud Incident

- The Crime: ATM Fraud
  - 100 ATM card numbers stolen & used

What’s the possible loss for the bank?

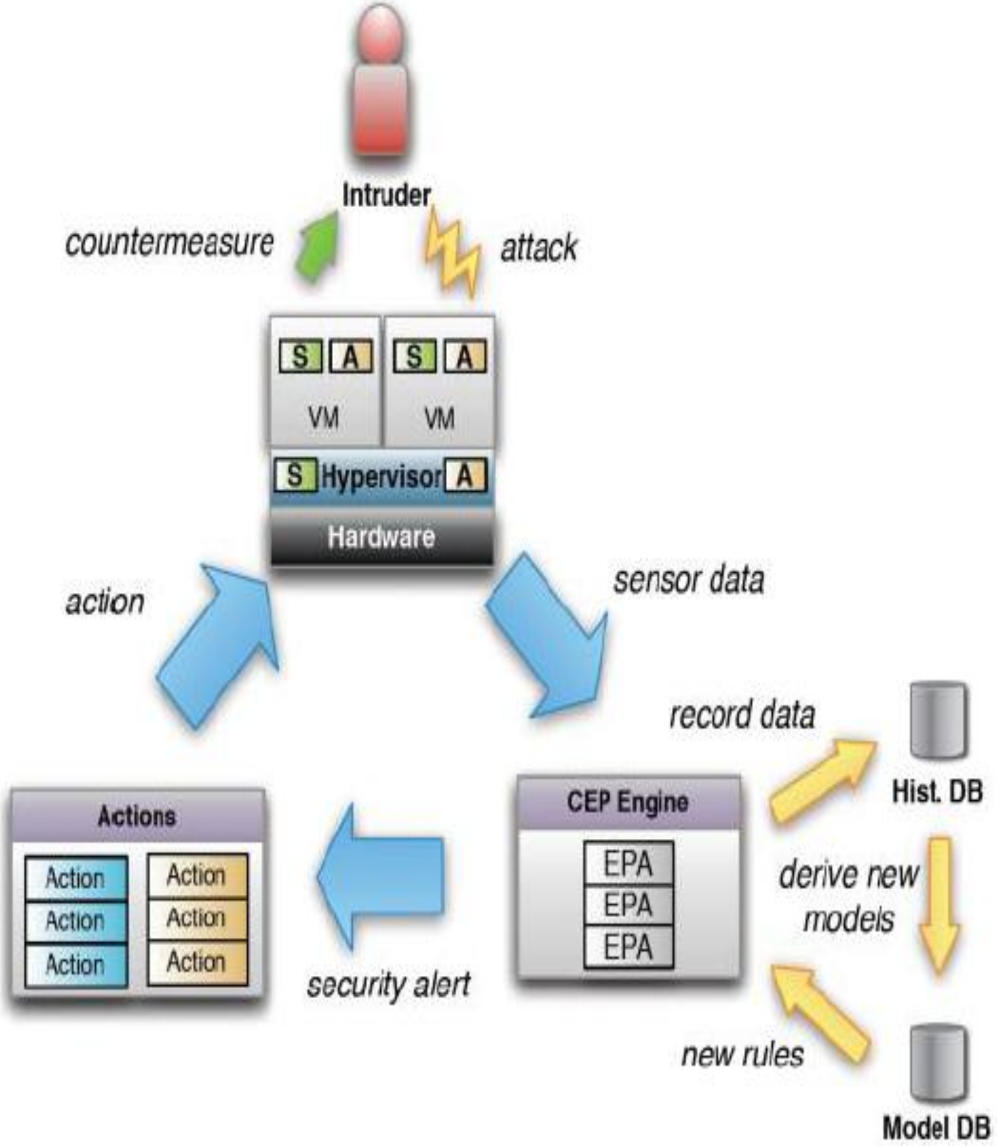
The Cost: $9,000,000

- 130 different ATM machines
- 49 cities worldwide
- 30 minutes

Use Case: Logistics

Description	**Fleet & Operation Management** Tracking fleet and cargo locations and meeting SLAs.
Challenges	• Overwhelming events (from warehouse, vehicle management & RFID systems, GPS devices, and environmental sensors) • No actionable insights to effectively manage resources • Non-optimal capacity usage
Objectives	• Instant detection of route deviations and updates to estimated time of arrivals (ETA) • Meet customer SLAs & arm them with info to make contingency plans • More effectively direct material, trucks, people, etc. to places where required

ACCEPT

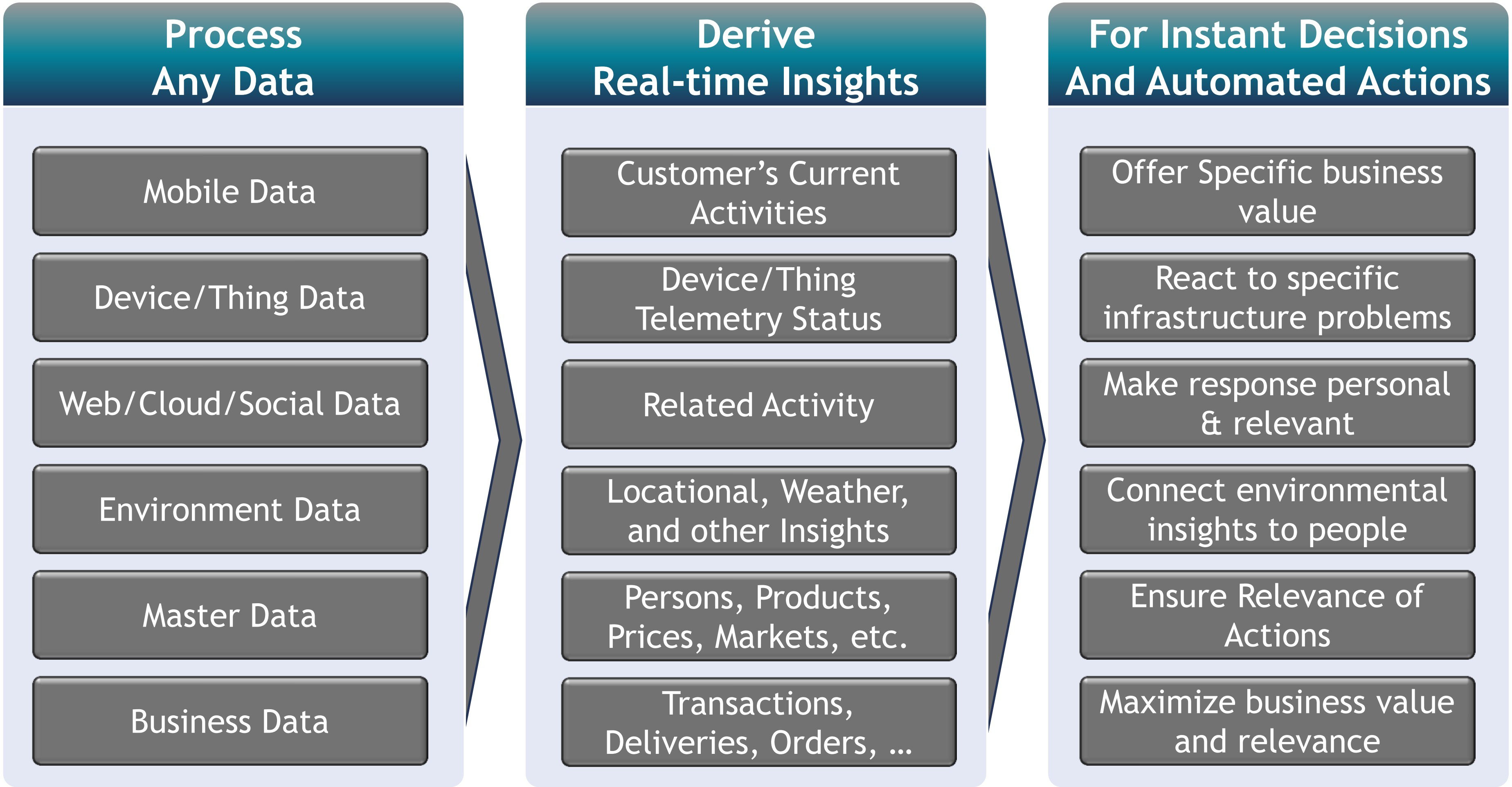


Recognition, analysis and handling of security related anomalies in virtualized computer systems

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

www.accept-projekt.de

software AG | controlware communicationssystems | Philipps Universität Marburg | TECHNISCHE UNIVERSITÄT DARMSTADT | Fraunhofer SIT



Process Any Data

- Mobile Data
- Device/Thing Data
- Web/Cloud/Social Data
- Environment Data
- Master Data
- Business Data

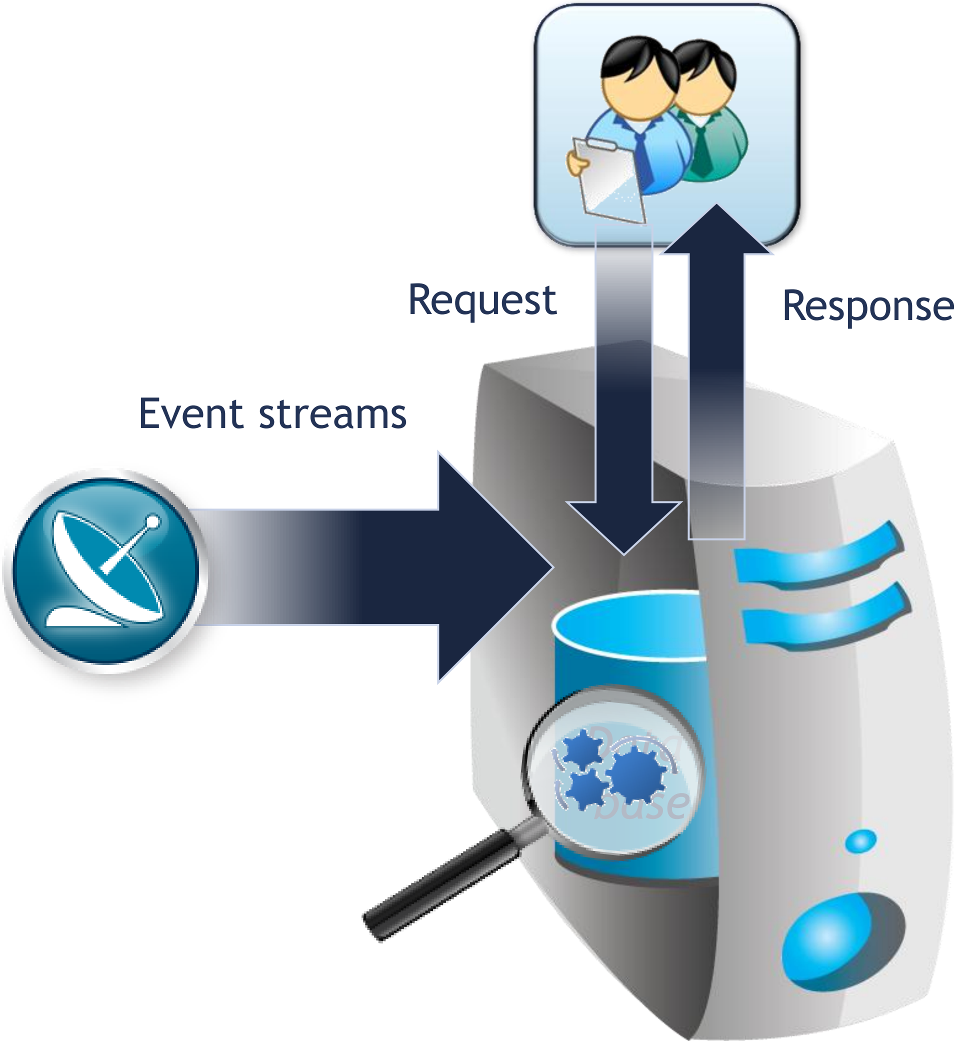
Derive Real-time Insights

- Customer's Current Activities
- Device/Thing Telemetry Status
- Related Activity
- Locational, Weather, and other Insights
- Persons, Products, Prices, Markets, etc.
- Transactions, Deliveries, Orders, …

For Instant Decisions And Automated Actions

- Offer Specific business value
- React to specific infrastructure problems
- Make response personal & relevant
- Connect environmental insights to people
- Ensure Relevance of Actions
- Maximize business value and relevance

Traditional „store-and-analyze“



Two phases:

1. Store data
2. Process one-time queries (pull-based analysis)

Problems

- Data store grows permanently

➔ Expensive search & analysis

- Not designed for continuous query evaluation

➔ Workarounds entail high load

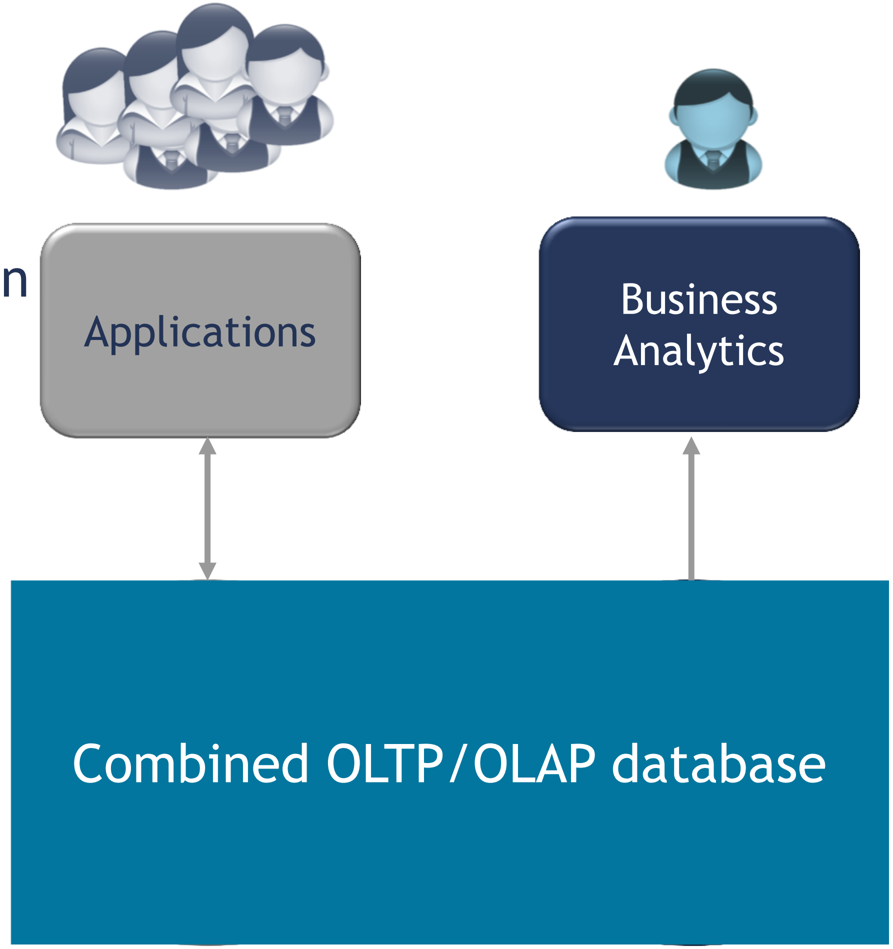
Observations on data streams

- Data stream sources often not collocated
- Data often quite granular
  - Typically no requirement to persist single values (continous temperature monitoring, vehicle position etc.)
  - Typically only data combinations indicate something relevant → classical aggregates, time series interpretation
- Structured and non-structured data
- Fast processing required
  - Batch-orientation not suitable

Source: http://de.wikipedia.org/wiki/Datei:Six-thermometer-disassembled26.jpg

Classical Data Management Architecture

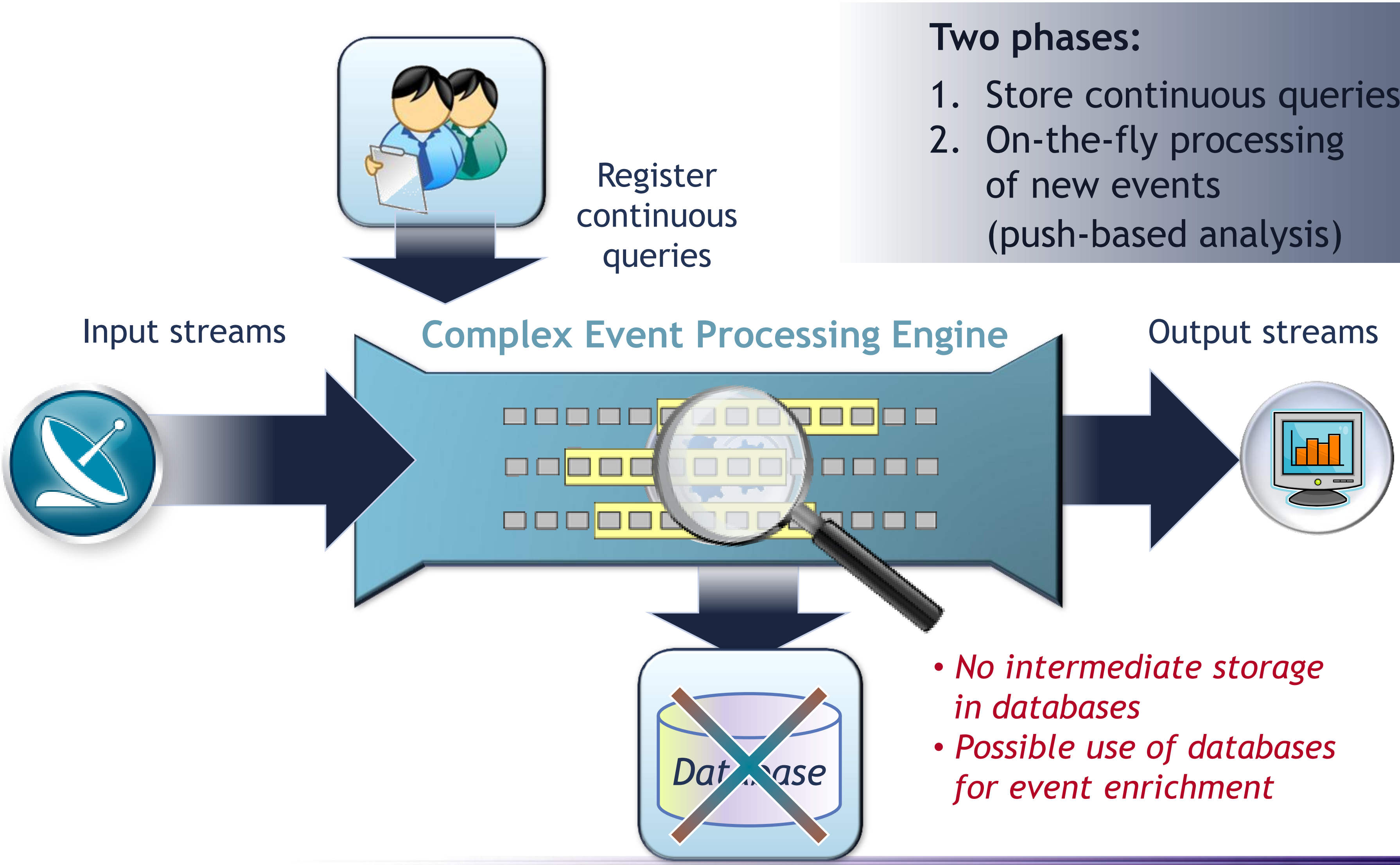
- Operational and analytical environments are separated
- Mainly based on structured data
- Data exchange and transformation
- Different ways end-user interacts with data
- Queries are triggered explictly



Event Stream Processing

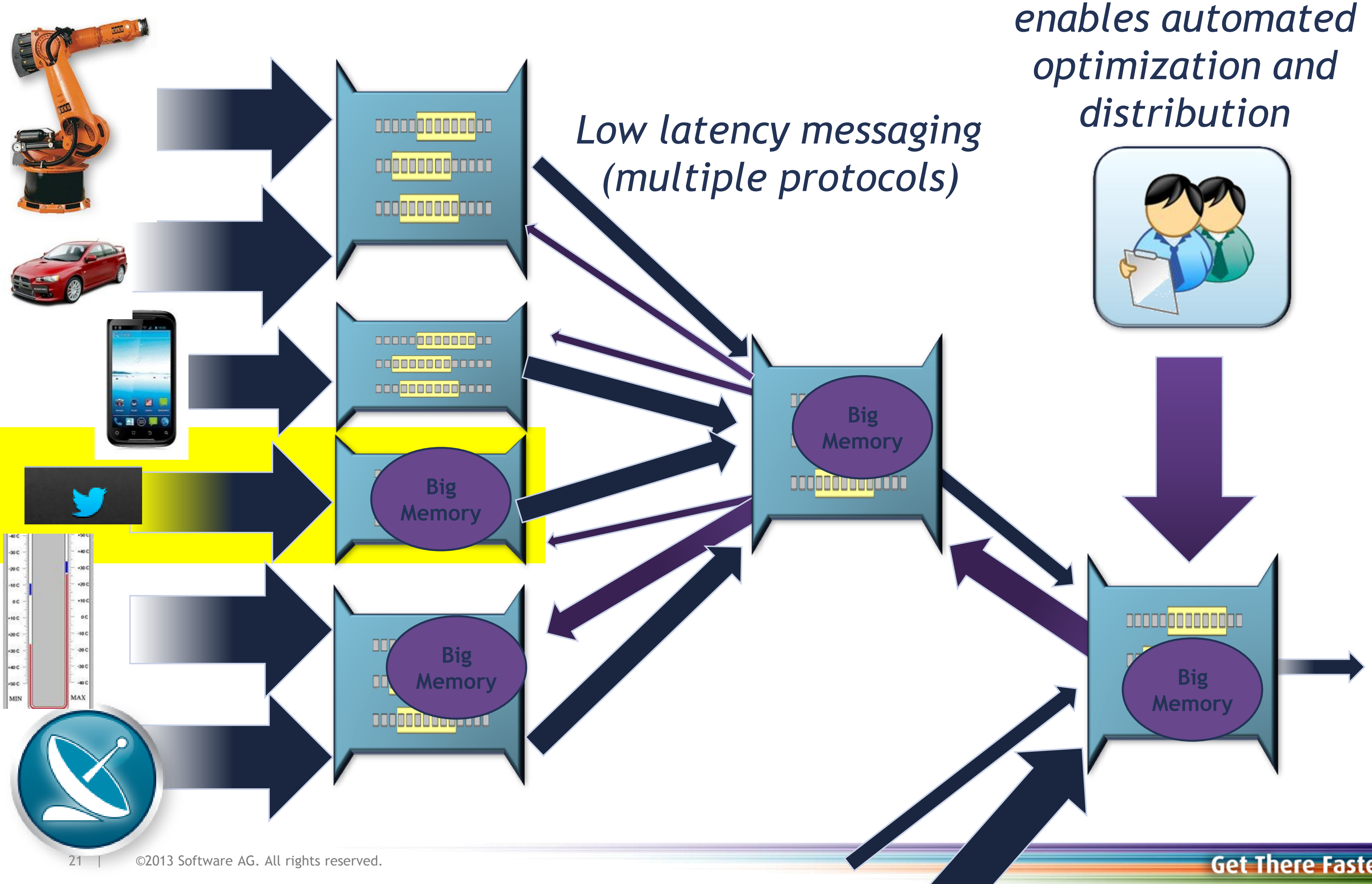
Two phases:

1. Store continuous queries
2. On-the-fly processing of new events (push-based analysis)



- *No intermediate storage in databases*
- *Possible use of databases for event enrichment*

Big Data: Process and forget



Low latency messaging (multiple protocols)

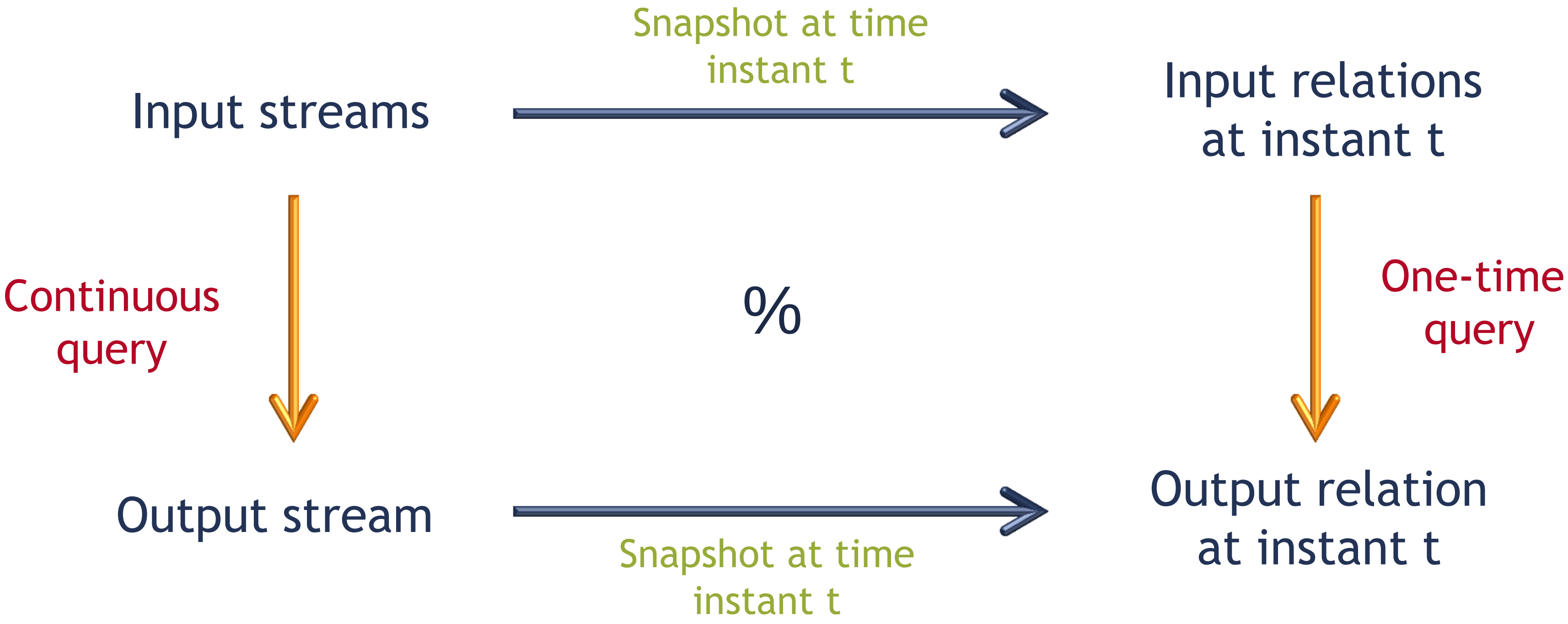
Descriptive language: enables automated optimization and distribution

Event processing language requirements

- SQL-like functionality
  - Filtering, grouping, aggregation, correlation
- Windowing (time, count, sliding)
- Pattern matching
- Non-event detection
- Enrichment
- Exact semantics
  - Predictable and repeatable
  - snapshot
- Optimizable

Semantic Compliance with Databases

- Exact specification of query results for any point in time
  - DBS would produce identical results if applied to every single time instant
- All conventional transformation rules applicable due to snapshot reducibility
  - ➔ Powerful query optimizations applicable



Snapshot at time instant t

Input streams

Input relations at instant t

Continuous query

%

One-time query

Output stream

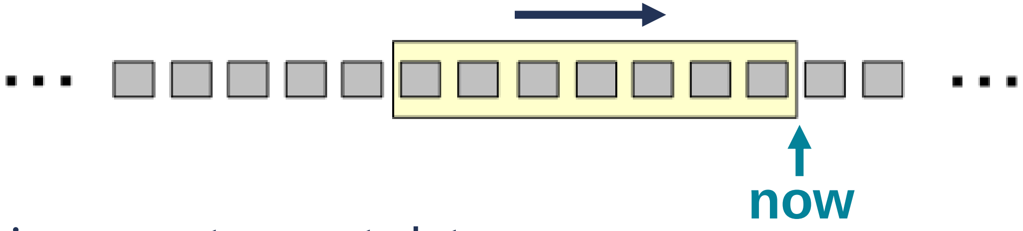
Output relation at instant t

Snapshot at time instant t

Sliding Windows

- Problem with some continuous queries
  - Computation of exact answer not possible
  - High-quality approximate answers are often acceptable

- Solution
  - Restriction of query range to finite sliding windows

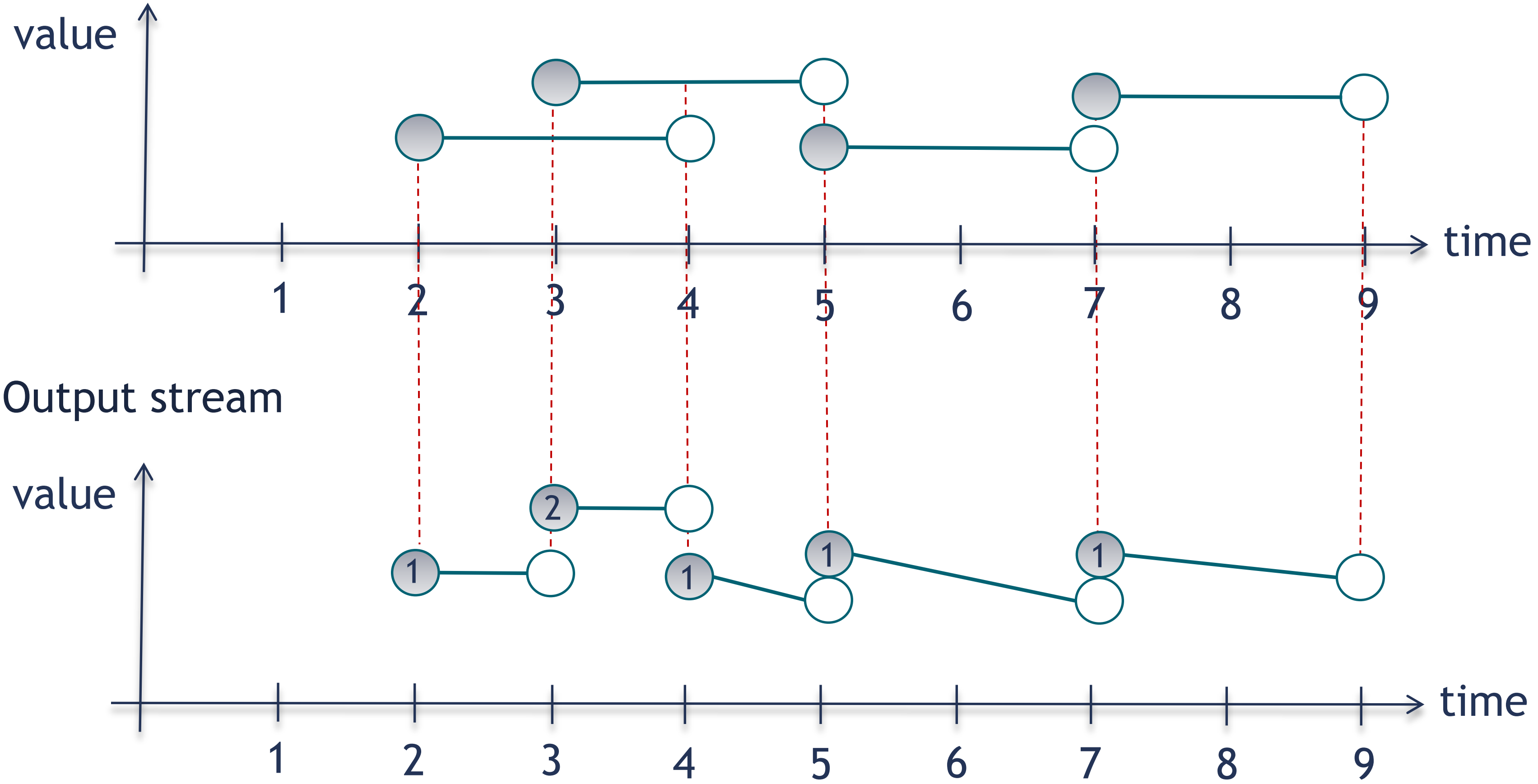


- Benefits
  - Emphasis on most recent data
    ➔ more important than older data
  - Query semantics can be defined precisely
    ➔ deterministic answers

Example

```
SELECT COUNT(*)
FROM S WINDOW(RANGE 2);
```

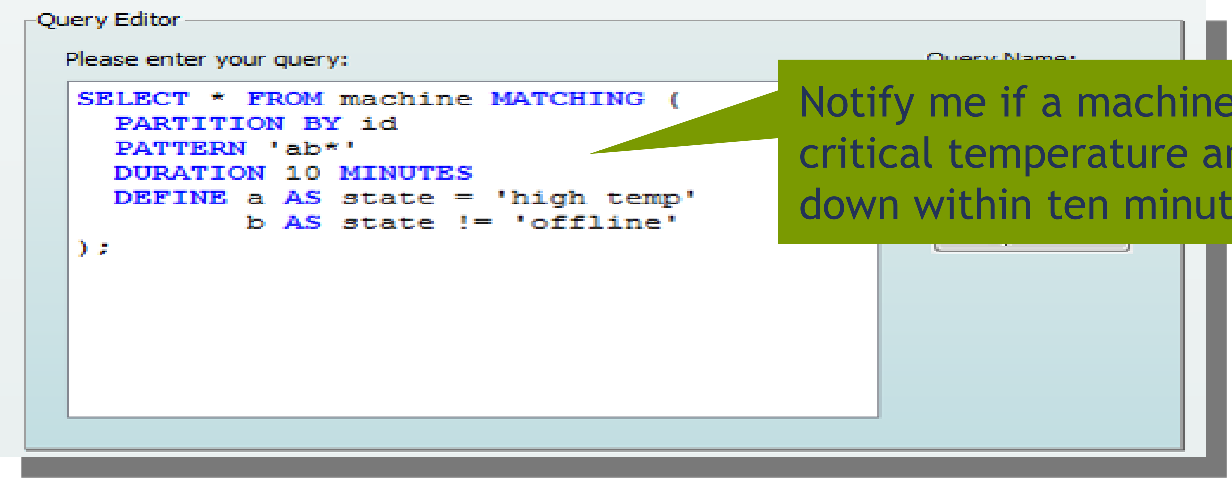
Input stream S



Output stream

Pattern Matching

- Detection of complex patterns
  - Pattern as sequence of events with certain conditions
  - Support of temporal patterns, state variables, set memberships, user-defined actions, etc.
  - Well-defined, deterministic results
  - Automaton-based implementation
- Easy usage: pattern specification in SQL queries



```
SELECT * FROM machine MATCHING (
   PARTITION BY id
   PATTERN 'ab*'
   DURATION 10 MINUTES
   DEFINE a AS state = 'high temp'
          b AS state != 'offline'
);
```

Notify me if a machine exceeds critical temperature and is not shut down within ten minutes!

Pattern Matching

- Motivation
  - Pattern queries are difficult to express in pure SQL
    - Joins can be used, but it isn't easy.
  - Determinism is important ➔ explanation of results

  - Examples
    - **Price Explosion Query**

      "Determine the *itemID* from items where the bid price increases rapidly."
    - **Stale Item Query**

      "Determine the *itemID* from items where no bid has arrived one minute after opening the auction."

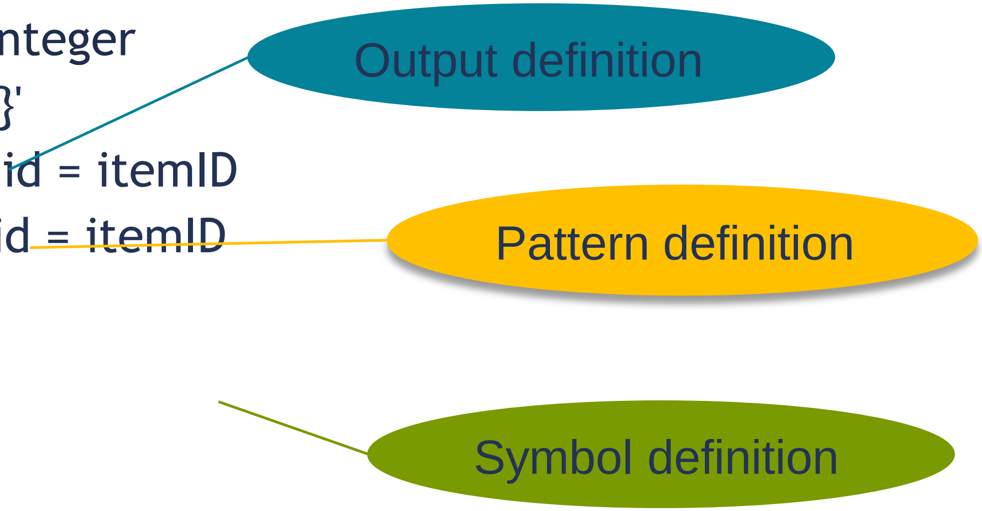
Pattern Matching

- Basic Idea
  - Detect sequence patterns in an event stream
- (Sequence) Pattern
  - Sequence patterns are described using regular expressions
    - a|b2 → {a,b,bb}
    - a|b* → {ε, a, b, bb, bbb, ...}
    - a|b+c → {a, bc, bbc, bbbc, ...}
  - Symbols can represent predicates (not only values)
    - Consideration of temporal constraints
      - "no bid b has arrived one minute after bid a"

Pattern Matching - Example 1

- “Determine the *itemID* from items with three subsequent bids in a row in the bid stream without intermediate bids on other items.”

-
```
 SELECT id
FROM Bid MATCHING (
         MEASURES id Integer
         PATTERN 'ab{2}'
         DEFINE    a DO id = itemID
                   b AS id = itemID
         );
```



Output definition

Pattern definition

Symbol definition

Pattern Matching - Example 2

- **Price Explosion Query**

“Determine the *itemID* from items where the bid price increases by more than 10% three times in a row.”

```
SELECT id AS itemID
FROM Bid MATCHING (
    PARTITION BY itemID
    MEASURES id Integer, currPrice Double
    PATTERN 'ab{3}'
    DEFINE a  DO id = itemID, currPrice = bid_price
                  b AS bid_price >= 1.1*currPrice DO currPrice = bid_price
);
```

Partitioning into substreams

Pattern Matching - Example 3

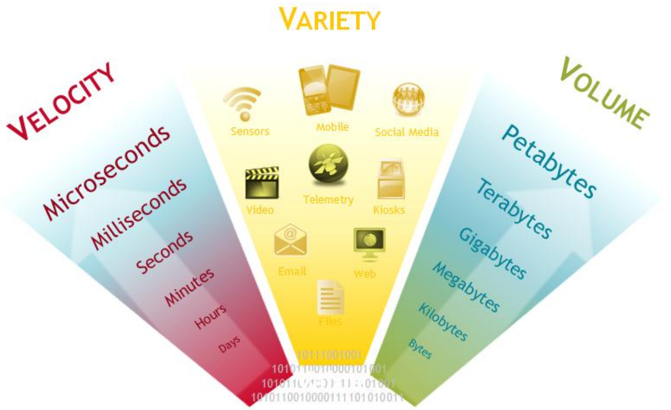
- **Stale Item Query**

“Determine the itemID from items where no bid has arrived within one minute after opening the auction.” (non-event detection)

```
SELECT *
FROM (SELECT itemID, 'open' AS action FROM OpenAuction
              UNION
           SELECT itemID, 'bid' AS action FROM Bid) AS openBidStream
MATCHING (
    MEASURES id Integer
    PATTERN 'ab*'
    DURATION 1 MINUTE
    DEFINE a AS action = 'open' DO id = itemID
                b AS itemID != id
);
```

Definition of the time interval

Summary



- Big Data is here and growing
- volume, velocity, variety, value

- Basic data are often not worth persisting
- also a matter of ecology

- Databases are only one building block in the picture

- Distributed multi-platform processing

Related topics

- Reliability
- Trust
- Privacy
  - National legislation
    - Fraud detection
    - Credit scoring
    - Customer profiling
  - Privacy-preserving data minig

Thank You!

Get There Faster