TIBCO™

Stream Processing

as Game Changer for the Internet of Things

Kai Wähner

kwaehner@tibco.com

@KaiWaehner

www.kai-waehner.de

LinkedIn / Xing → Please connect!

Key Messages

- Streaming Analytics processes Data while it is in Motion!
- Automation and Proactive Human Interaction are BOTH needed!
- Time to Market is the Key Requirement for most Use Cases!

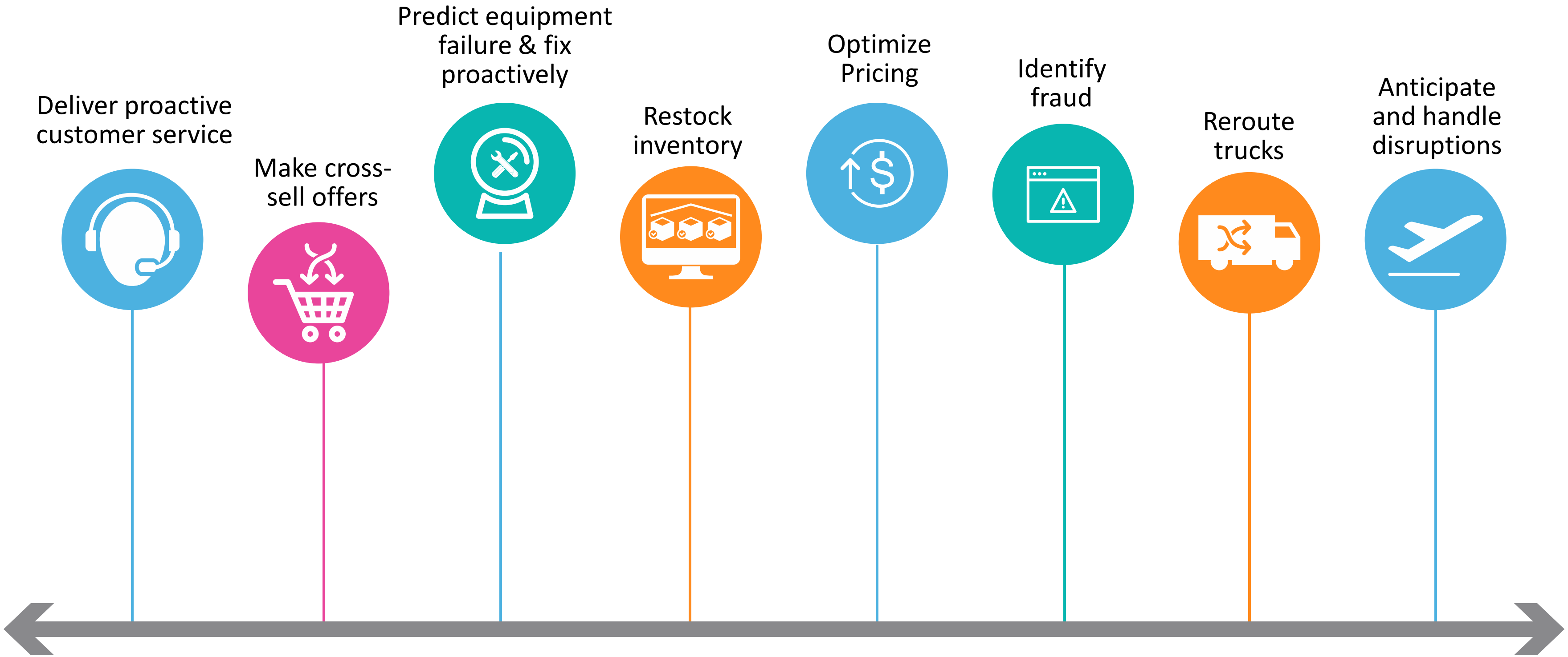
Agenda

- Real World Use Cases
- Introduction to Stream Processing
- Market Overview
- Relation to other Big Data Components
- Live Demo

TIBCO | Agenda

- **Real World Use Cases**
- Introduction to Stream Processing
- Market Overview
- Relation to other Big Data Components
- Live Demo

Find and Act on "Critical Business Moments"



"Business Moments" occur in Every Facet of Enterprise Operations, they drive competitive differentiation, customer satisfaction and business success!

Predictive Fault Management

“An outage on one well can cost $10M per hour. We have 20-100 outages per year.“

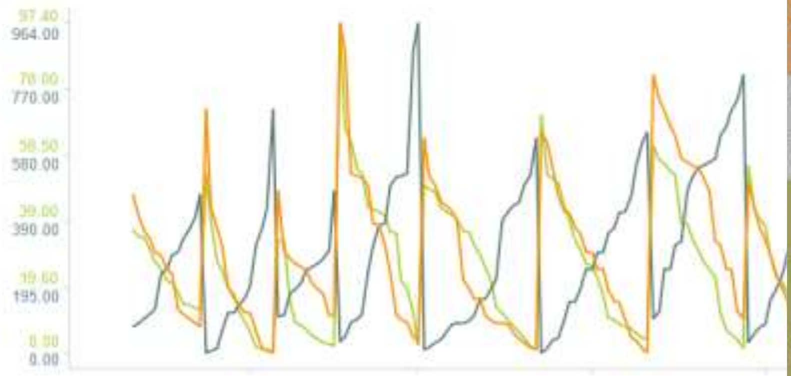
- Drilling operations VP, major oil company

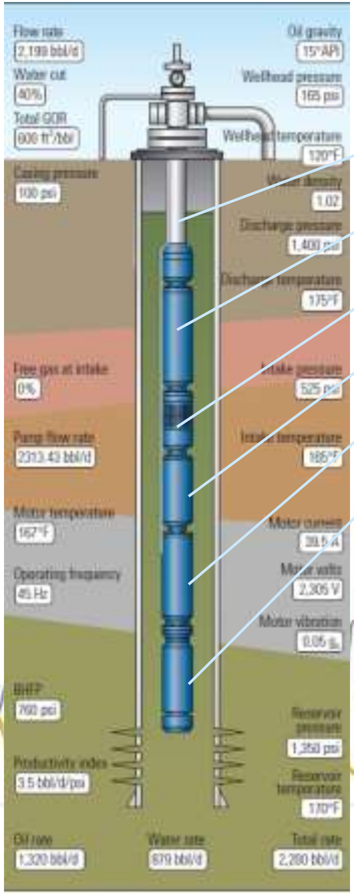
TIBCO Predictive Analytics (Fault Management)

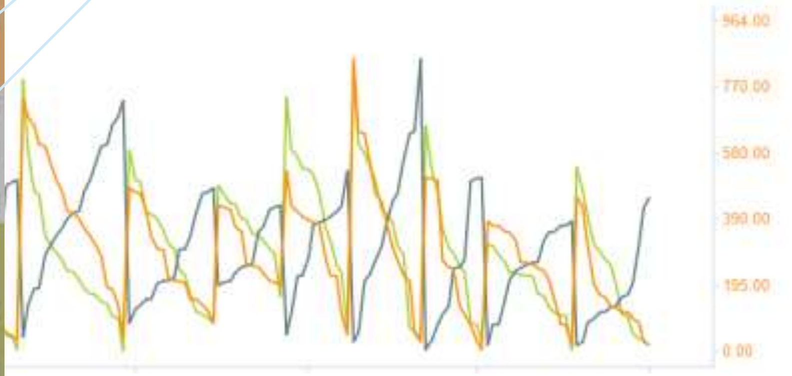
Data Monitoring

- Motor temperature
- Motor vibration
- Current
- Intake pressure
- Intake temperature

➢ Flow





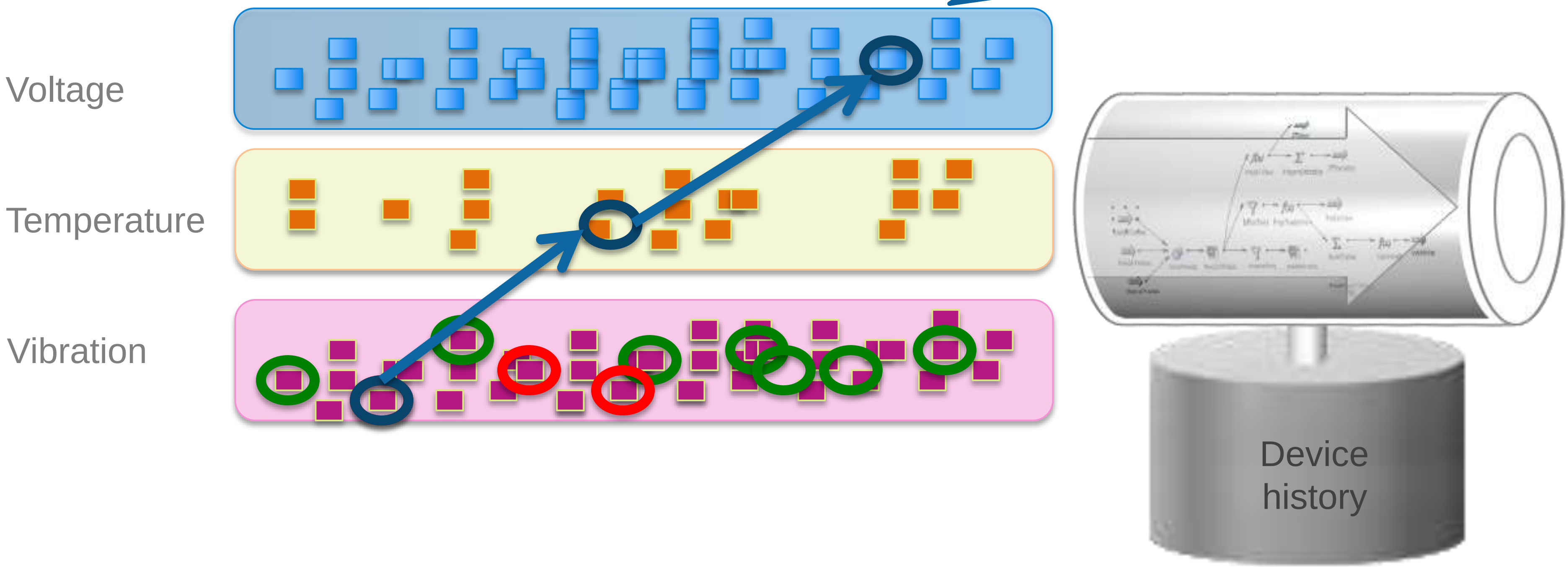


Electric Submersible Pumps (ESP)

Electrical power cable
Pump
Intake
Protector
ESP motor
Pump monitoring unit

Predictive Analytics (Fault Management)

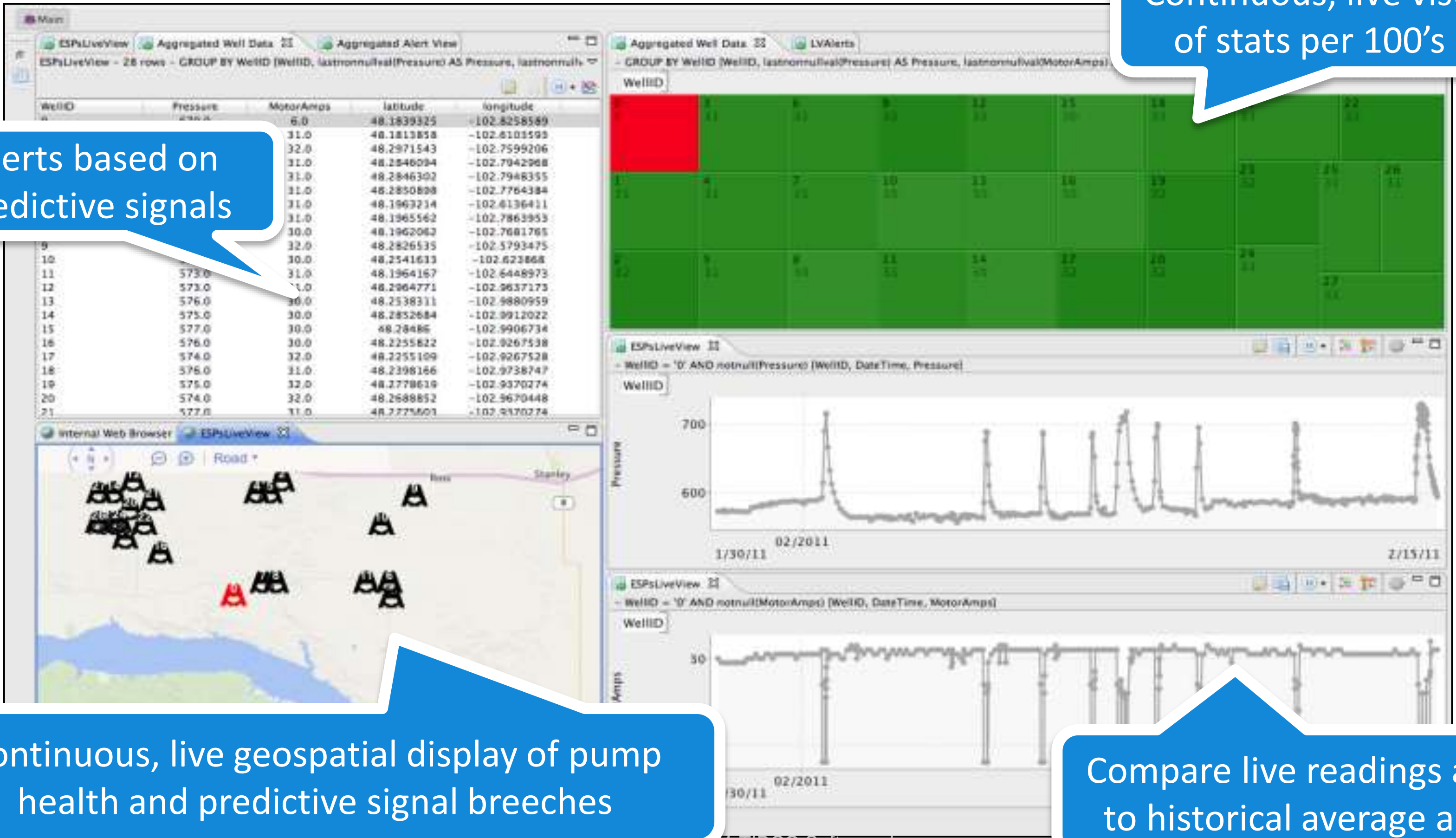
<u>Temporal analytic</u>: "If **vibration spike** is followed by **temp spike** then **voltage spike** [within 12 minutes] then flag **high severity alert**."



Live Surveillance of Equipment

Continuous, live visualization of stats per 100's of wells

Alerts based on predictive signals



Continuous, live geospatial display of pump health and predictive signal breeches

Compare live readings and signals to historical average and means

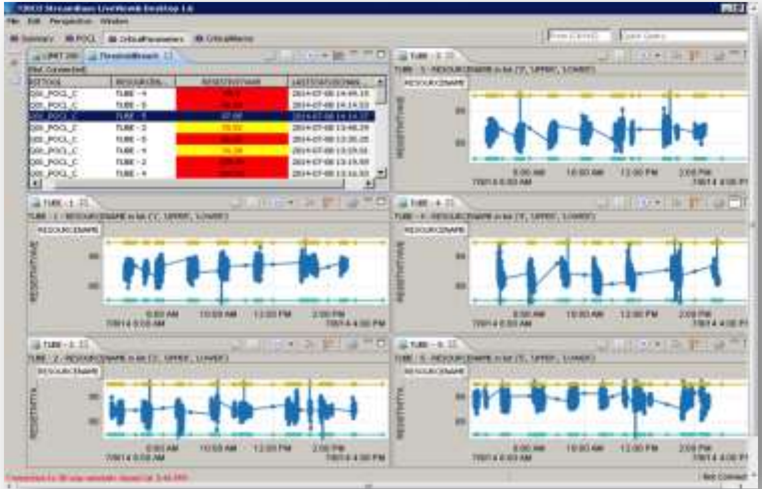
Smart Manufacturing

““For every 1% increase in shipped product, we make $11MM in profit. The demand is there, we just need to fulfill it.“

- Head of Quality, Solar Panel Manufacturer

IoT for High Tech Manufacturing Yield Optimization

- Before: Solar Panel Manufacturer with No Unified View of Manufacturing Process
  - Multiple manufacturing facilities, multiple processes – no way to compare production to yield expectations
- Negative Consequences: Sub-Optimal Production
  - Operations are sub-optimal: high tolerance leads to better yield but less output; tight tolerance means high throughput but lower yield
- Business Outcome: Higher Yield and More Runs
  - Process Manufacturing can run tighter tolerances and adjust them mid-run, predicting yield and adjusting to changing variables
  - Systems proactively re-route high-value customers around affected network areas in real-time
- How We Do It: The TIBCO Fast Data Platform
  - IoT, Spotfire, StreamBase, and TERR for predictive modeling, high-speed network by TIBCO



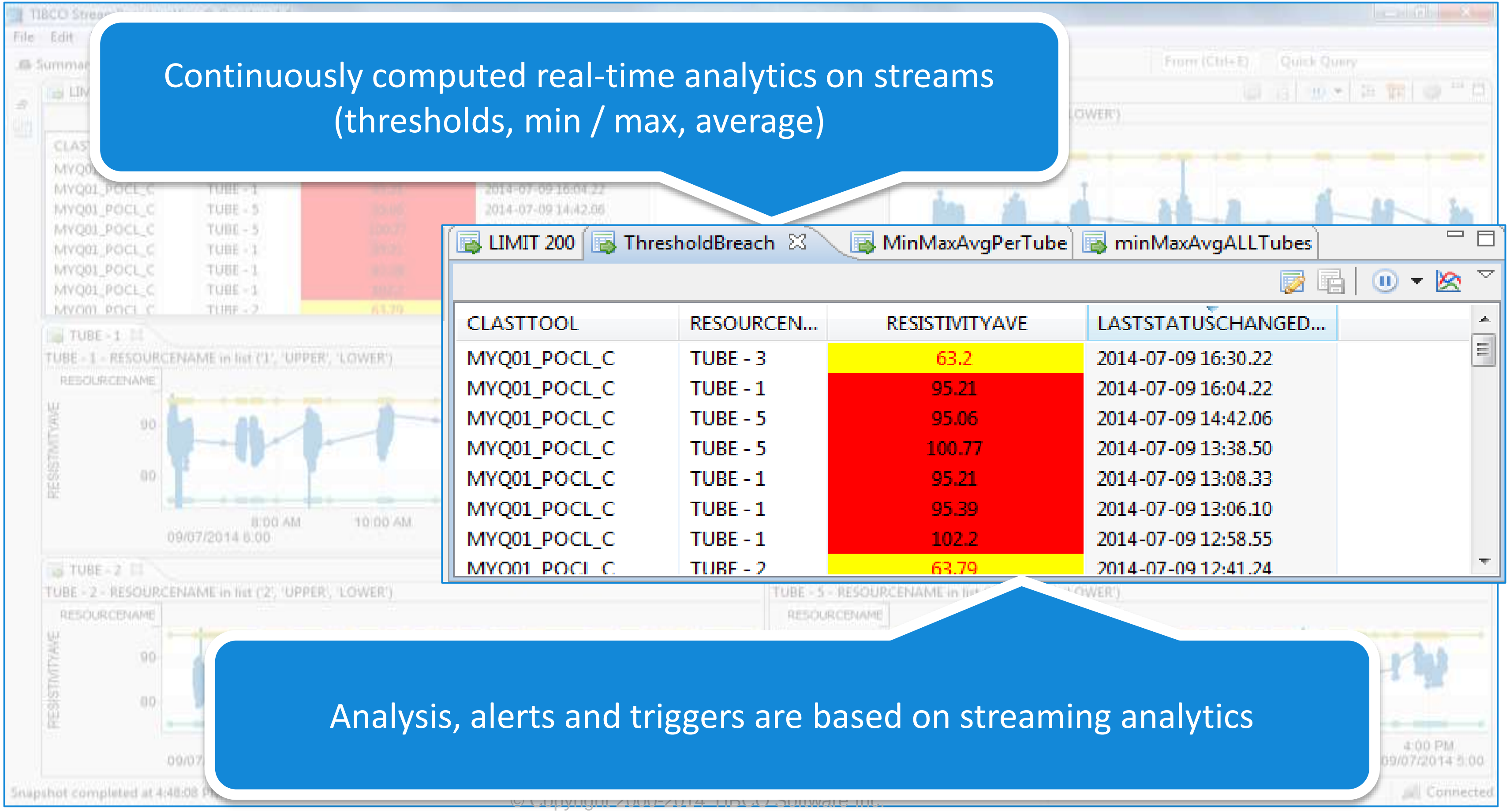
Real Time

"For every 1% increase in shipped product, we make $11MM in profit. The demand is there, we just need to fulfill it."

- Head of Quality, Solar Panel Manufacturer

TIBCO | High Tech Manufacturing Yield Optimization

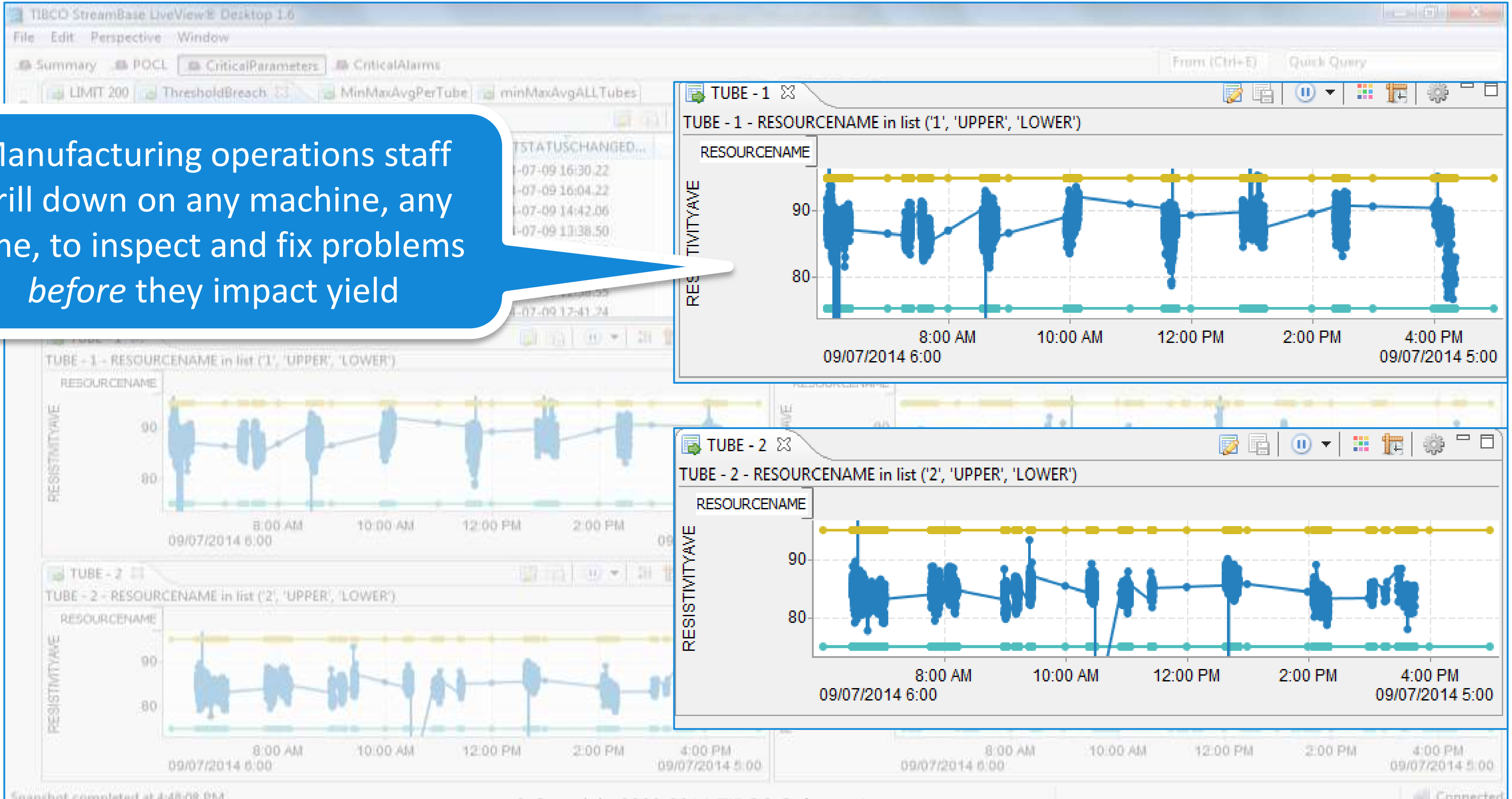
Continuously computed real-time analytics on streams (thresholds, min / max, average)



Analysis, alerts and triggers are based on streaming analytics

High Tech Manufacturing Yield Optimization

Manufacturing operations staff drill down on any machine, any time, to inspect and fix problems *before* they impact yield

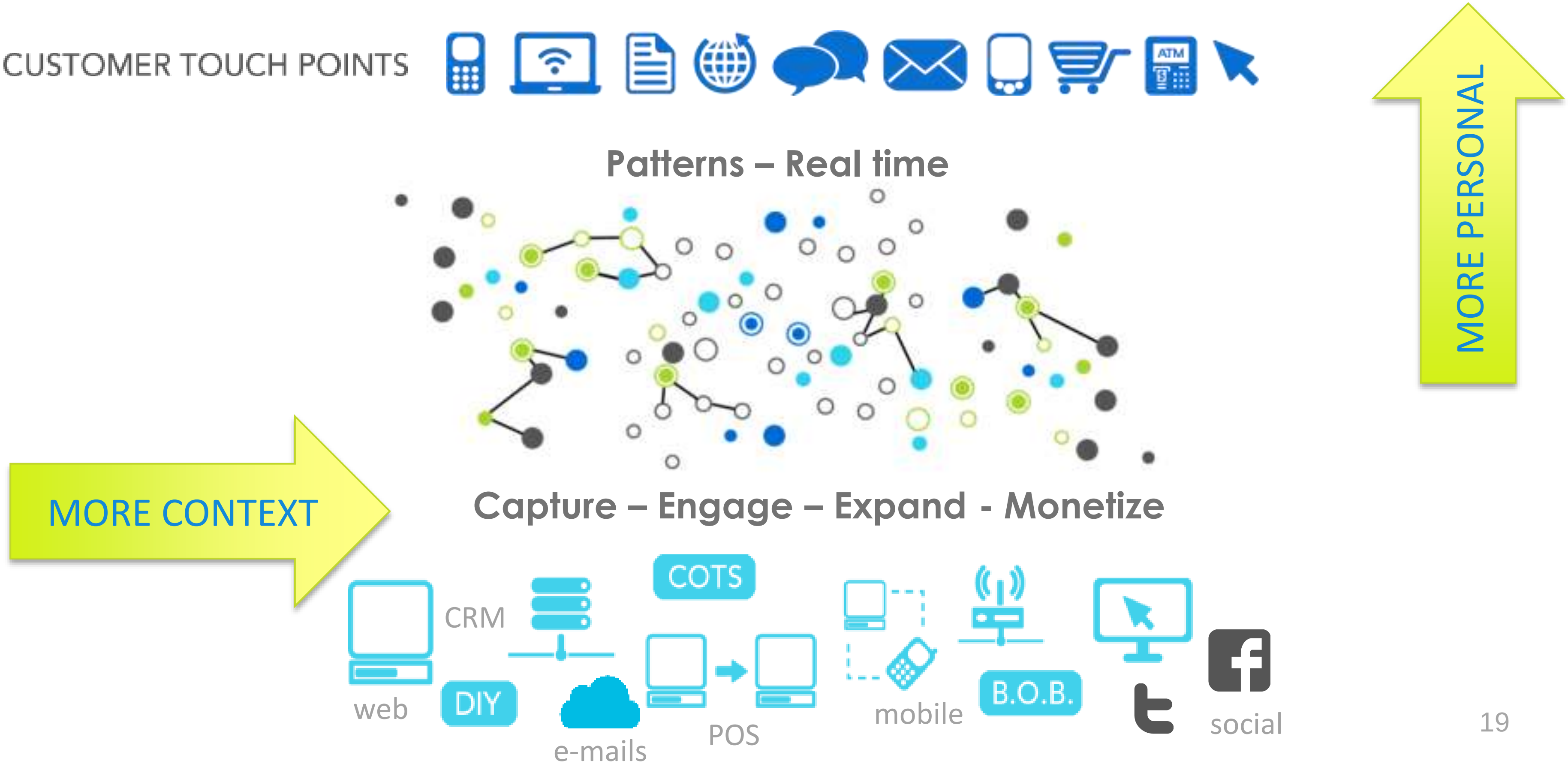


Crowd Management

“Turn the customer into a fan and increase revenue significantly.“

TIBCO™ | Sacramento Kings → World's Smartest Building

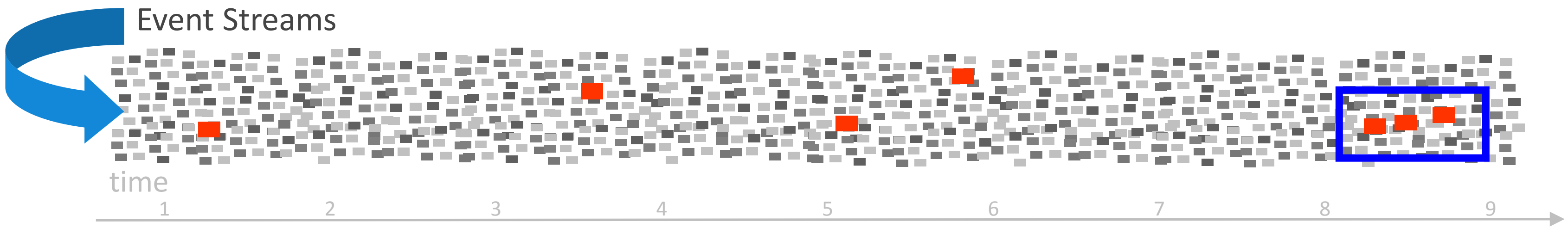
All Customers are different… Treat them that way…



Great success stories, but …

… how to realize these use cases?

- Real World Use Cases
- **Introduction to Stream Processing**
- Market Overview
- Relation to other Big Data Components
- Live Demo

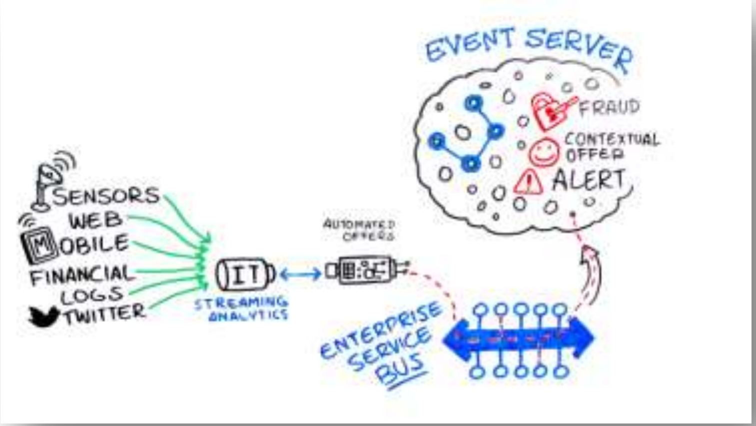


- Continuous Queries
- Sliding Windows
- Filter
- Aggregation
- Correlation
- …

Operational Intelligence in Action

Machine-to-Machine Automation

Automated action based on models of history combined with live context and business rules



The Challenge:

<u>Create</u>, *<u>understand</u>*, and *<u>deploy</u>* algorithms & rules that automate key business reactions

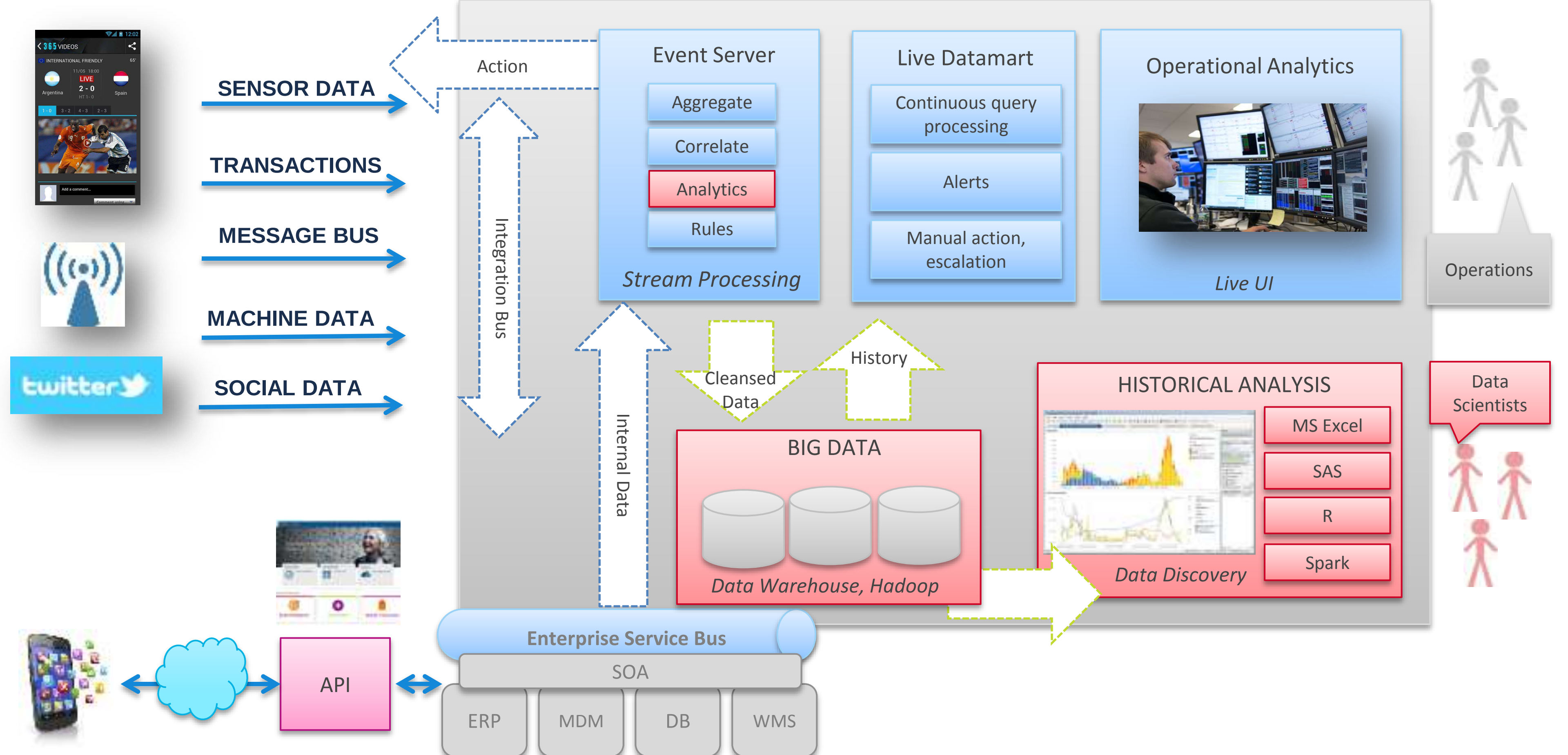
Actions by Operations

Human decisions in real time informed by up to date information

The Challenge:

Empower operations staff to *<u>see and seize key business moments</u>*

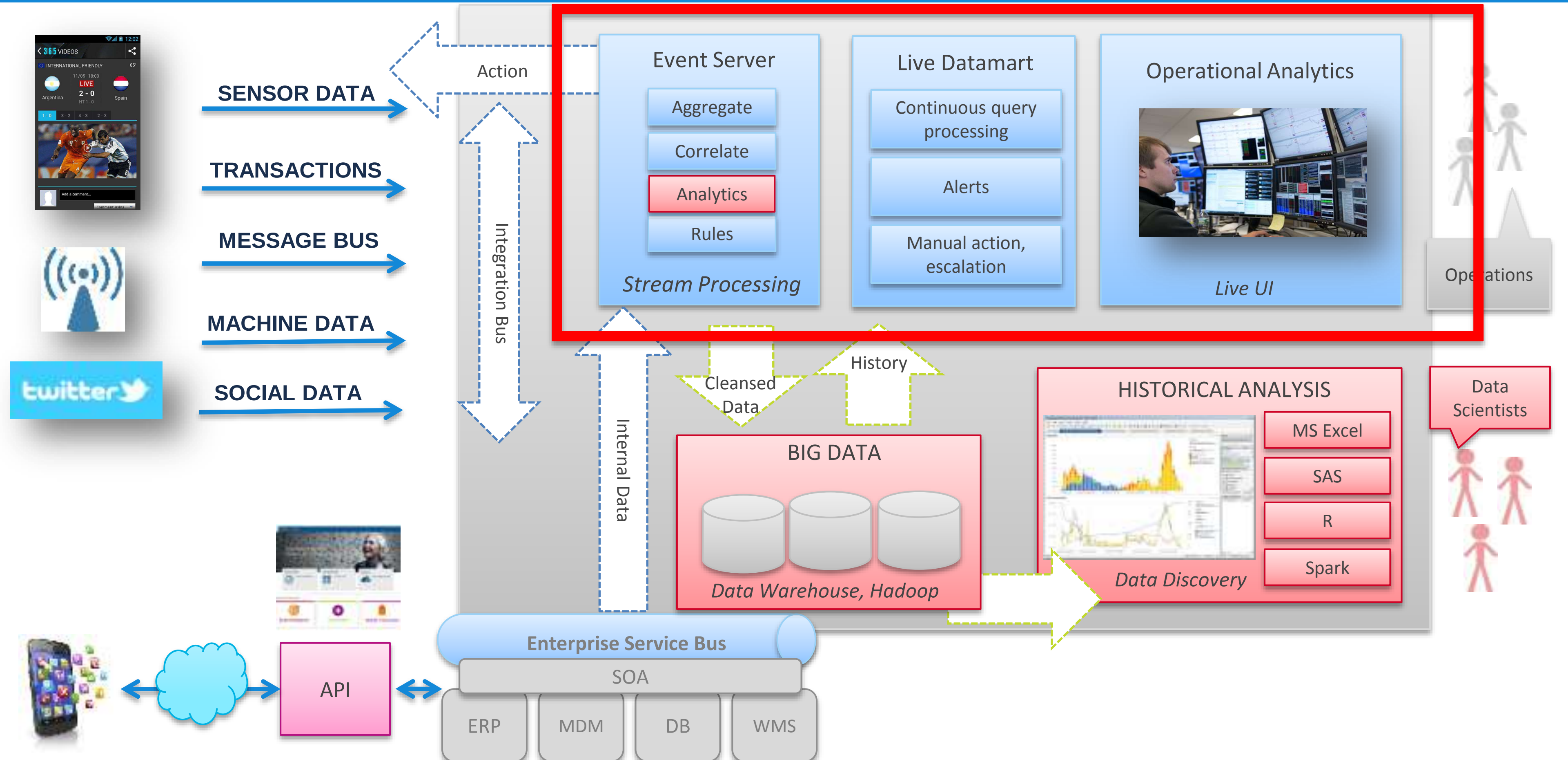
Streaming Analytics Reference Architecture



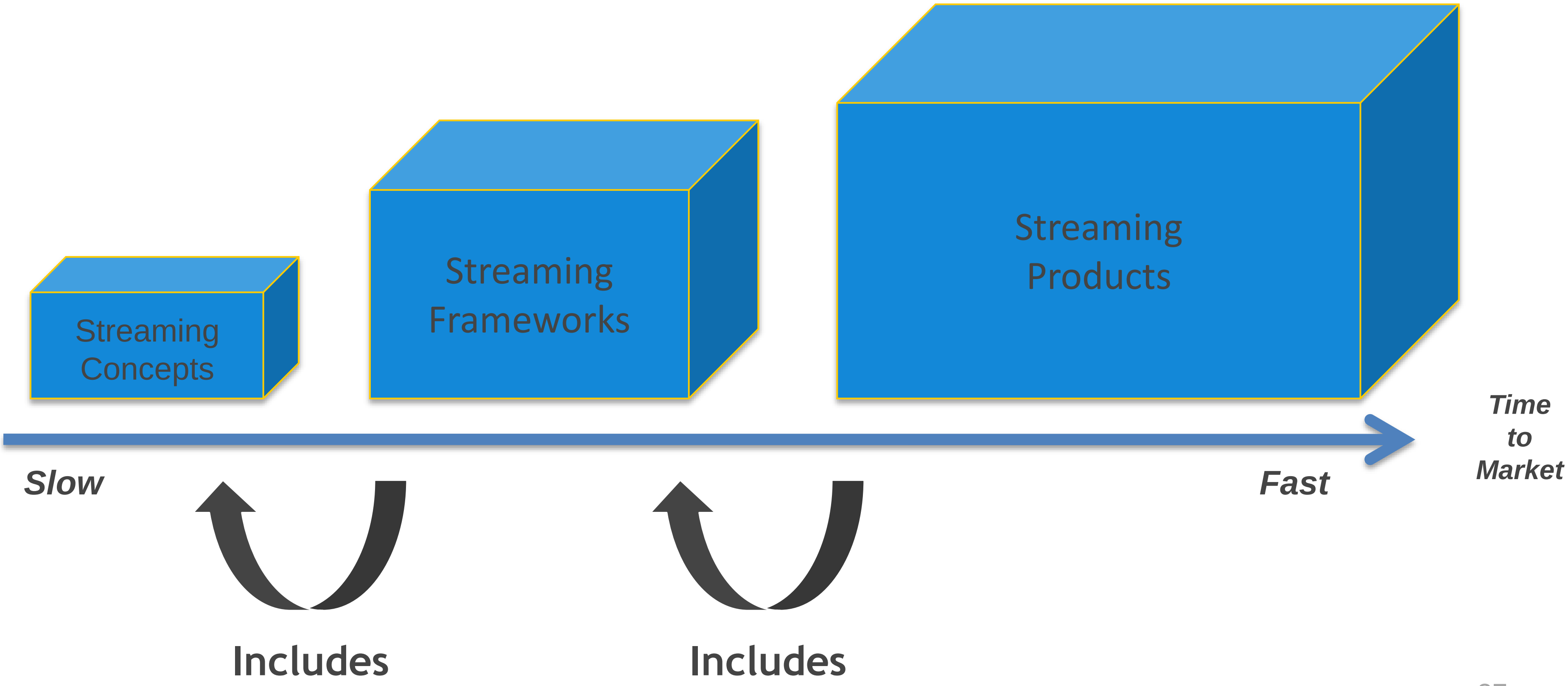
Agenda

- Real World Use Cases
- Introduction to Stream Processing
- **Market Overview**
- Relation to other Big Data Components
- Live Demo

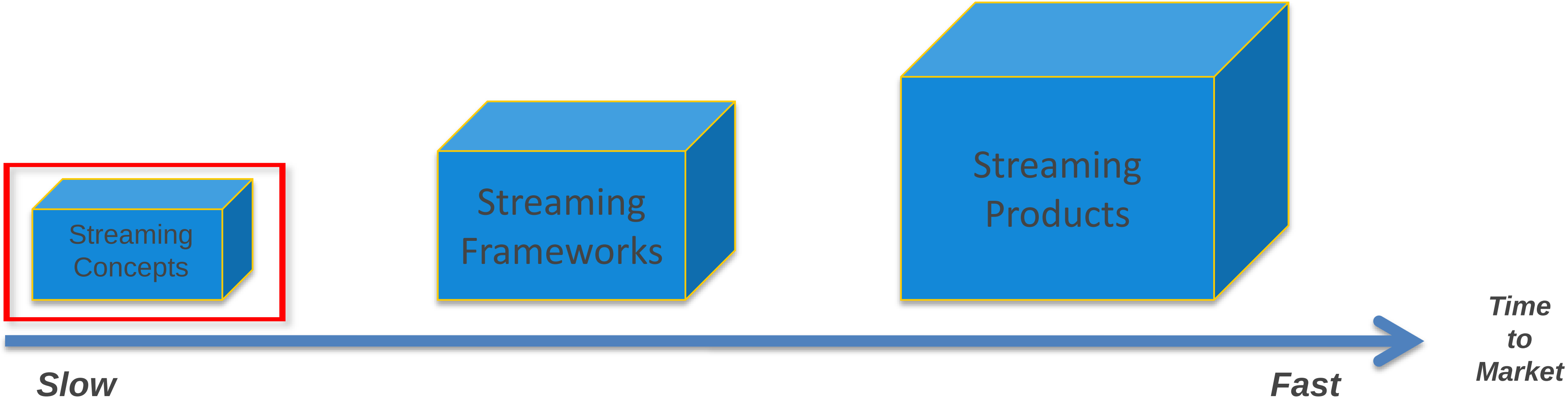
TIBCO | Streaming Analytics Reference Architecture



Alternatives for Stream Processing



TIBCO | What Streaming Alternative do you need?



Concepts (Continuous Queries, Sliding Windows)
Patterns (Counting, Sequencing, Tracking, Trends)

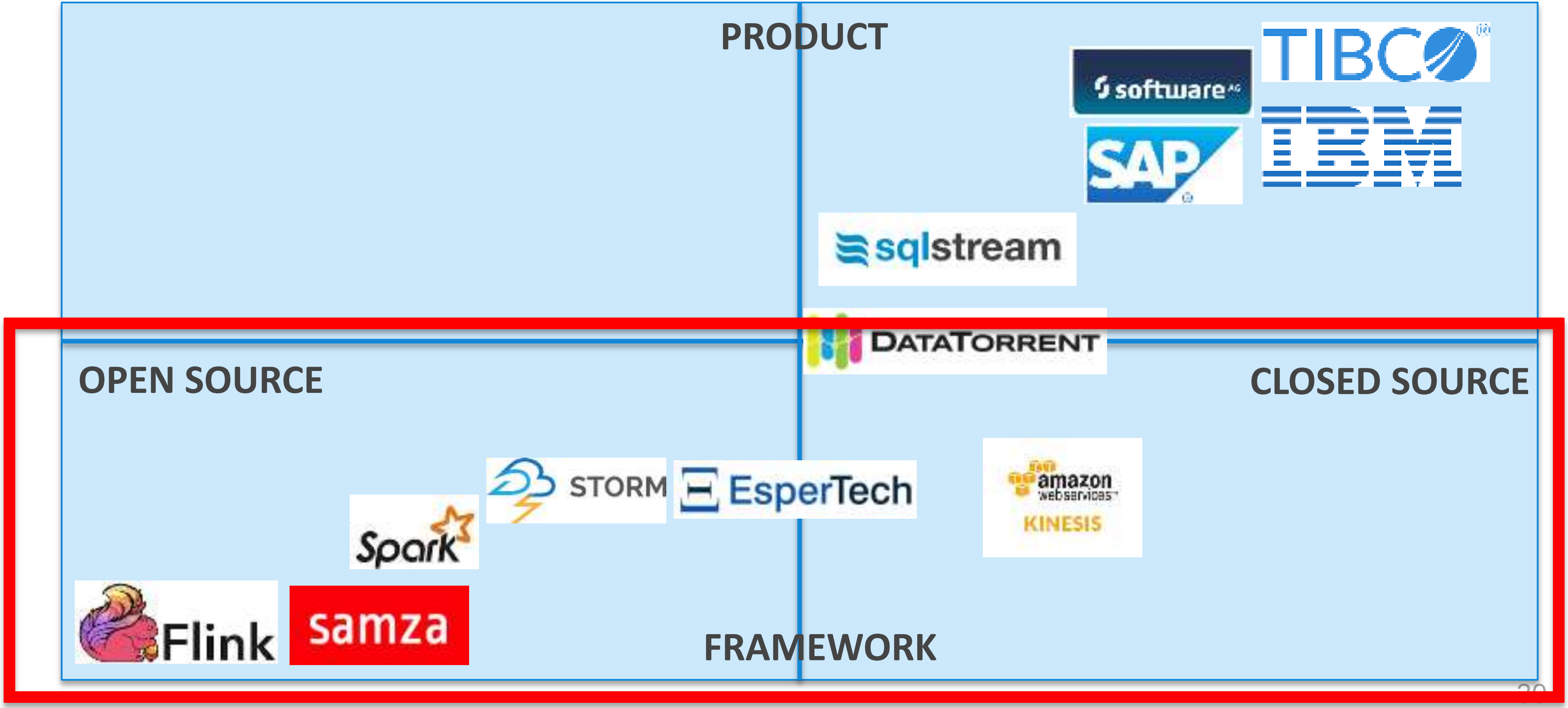
Build everything by yourself! ☹

Usually not an option ...

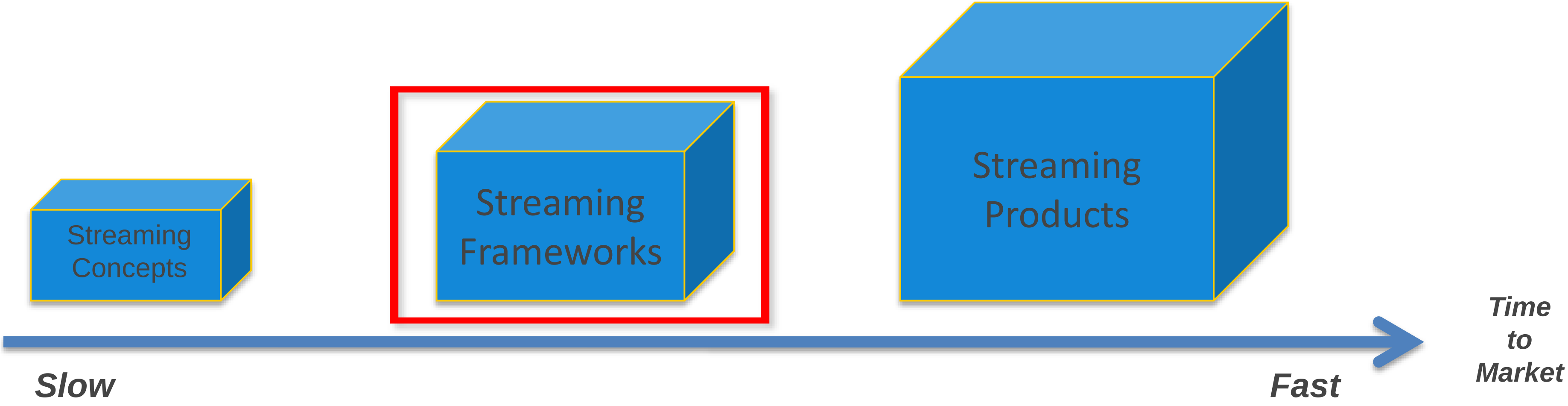
… as there are a lot of Frameworks and Products available!

TIBCO | Alternatives

(no complete list!)



What Streaming Alternative do you need?

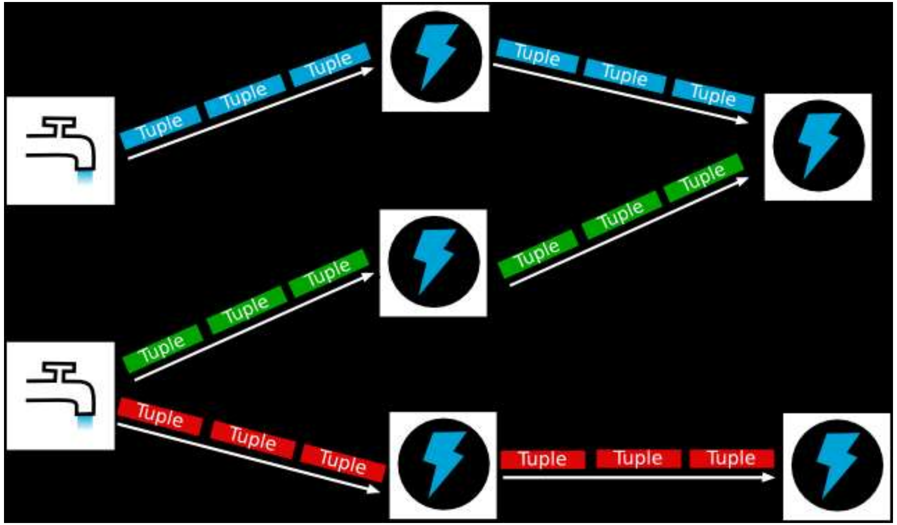


Library (Java, .NET, Python)
Query Language (often similar to SQL)
Scalability (horizontal and vertical, fail over)
Connectivity (technologies, markets, products)
Operators (Filter, Sort, Aggregate)

Apache Storm

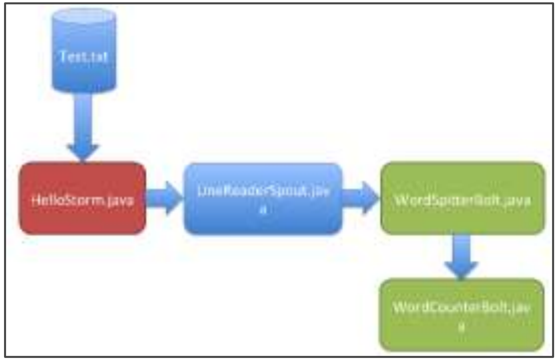
```
nathanmarz says: Storm is a distributed realtime computation system. Similar to how
hadoop provides a set of general primitives for doing batch processing. Storm provides
a set of general primitives for doing realtime computation. Storm is simple, can be used
with any programming language, and is a lot of fun to use!
```

Spout



Bolt

TIBCO | Apache Storm – Hello World




```java
public static void main(String[] args) throws Exception{
        Config config = new Config();
        config.put("inputFile", args[0]);
        config.setDebug(true);
        config.put(Config.TOPOLOGY_MAX_SPOUT_PENDING, 1);

        TopologyBuilder builder = new TopologyBuilder();
        builder.setSpout("line-reader-spout", new LineReaderSpout());
        builder.setBolt("word-spitter", new WordSpitterBolt()).shuffleGrouping("line-reader-spout");
        builder.setBolt("word-counter", new WordCounterBolt()).shuffleGrouping("word-spitter");

        LocalCluster cluster = new LocalCluster();
        cluster.submitTopology("HelloStorm", config, builder.createTopology());
        Thread.sleep(10000);

        cluster.shutdown();
}
```

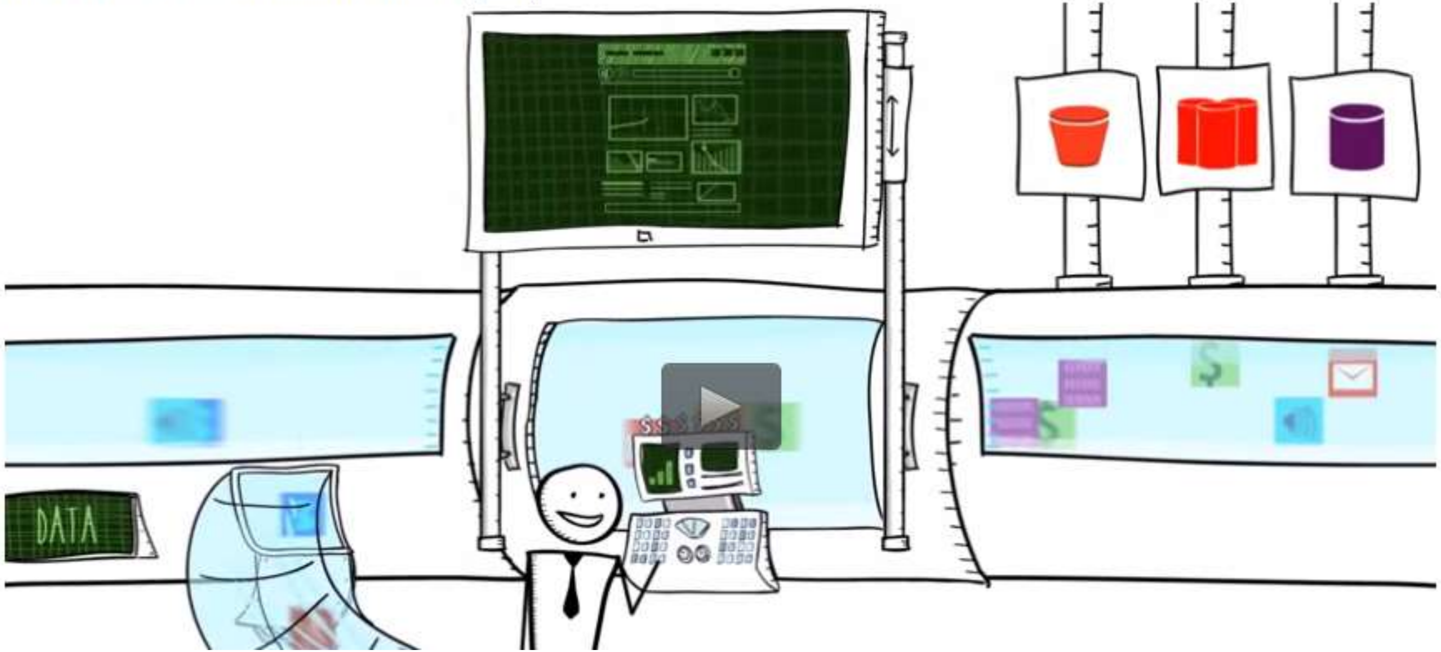
```java
public class WordCounterBolt implements IRichBolt{
        Map<String, Integer> counters;
        private OutputCollector collector;
        @Override
        public void prepare(Map stormConf, TopologyContext context,
                        OutputCollector collector) {
                this.counters = new HashMap<String, Integer>();
                this.collector = collector;
        }
        @Override
        public void execute(Tuple input) {
                String str = input.getString(0);
                if(!counters.containsKey(str)){
                        counters.put(str, 1);
                }else{
                        Integer c = counters.get(str) +1;
                        counters.put(str, c);
                }
                collector.ack(input);
        }
        @Override
        public void cleanup() {
                for(Map.Entry<String, Integer> entry:counters.entrySet()){
                        System.out.println(entry.getKey()+" : " + entry.getValue());
                }
        }
        @Override
        public void declareOutputFields(OutputFieldsDeclarer declarer) {
        }

        @Override
        public Map<String, Object> getComponentConfiguration() {
                return null;
        }
}
```

http://wpcertification.blogspot.ch/2014/02/helloworld-apache-storm-word-counter.html

Amazon Kinesis

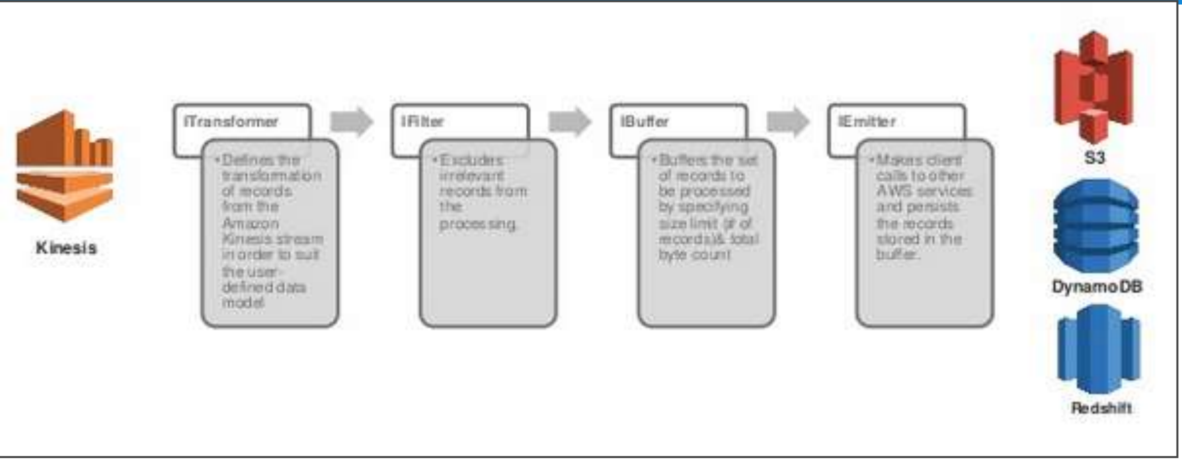
AWS S3 RedShift DynamoDB

Introduction to Amazon Kinesis (2:08)

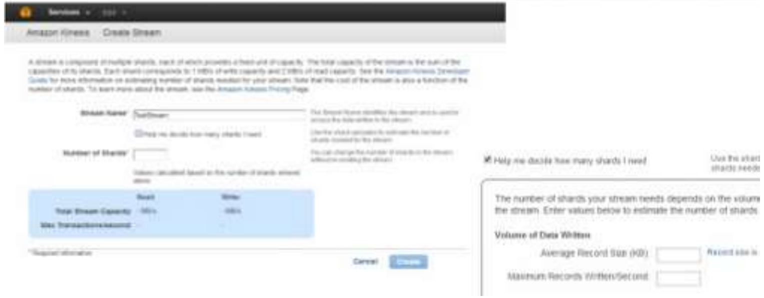


https://aws.amazon.com/kinesis/

Amazon Kinesis – Hello World



Creating and Sizing a Kinesis Stream




```
IRecordProcessorFactory recordProcessorFactory = new
SampleRecordProcessorFactory();
Worker worker = new Worker(recordProcessorFactory,
kinesisClientLibConfiguration);

int exitCode = 0;
try {
    worker.run();
} catch (Throwable t) {
    LOG.error("Caught throwable while processing data.", t);
    exitCode = 1;
}
```

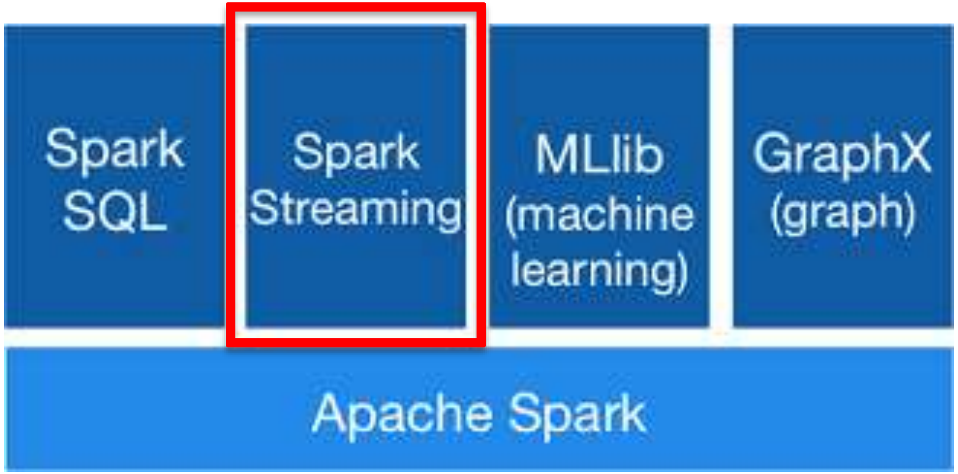
… is easy to setup and scale!

But you do not have full control ☹

Amazon Kinesis: Is It the Next Big Real-Time Data Processing Solution?

- Any data that is older than 24 hours is automatically deleted
- Every Kinesis application consists of just one procedure, so you can't use Kinesis to perform complex stream processing unless you connect multiple applications
- Kinesis can only support a maximum size of 50KB for each data item

http://diamondstream.com/amazon-kinesis-big-real-time-data-processing-solution/
(blog post from 2014, might be outdated, but shows that you do not have full control over a cloud service)



General Data-processing Framework

→ However, focus is especially on Analytics (these days)

http://fortune.com/2015/09/09/cloudera-spark-mapreduce/

TIBCO | Apache Spark – Focus on Analytics

Spark

Is Apache Spark going to replace Hadoop?

Published: 20, March 2015 07:33 am by Jameel Mohammed

Why Cloudera is saying 'Goodbye, MapReduce' and 'Hello, Spark'

by Derrick Harris @derrickharris SEPTEMBER 9, 2015, 7:06 AM EDT

Using Spark to Ignite Data Analytics

by eBay Global Data Infrastructure Analytics Team on 05/28/2014
in Data Infrastructure and Services, Machine Learning, Open Source

At eBay we want our customers to have the best experience possible. We use data analytics to improve user experiences, provide relevant offers, optimize performance, and create many, many other kinds of value. One way eBay supports this value creation is by utilizing data processing frameworks that enable, accelerate, or simplify data analytics. One such framework is Apache Spark. This post describes how Apache Spark fits into eBay's Analytic Data Infrastructure.

Forbes / Tech

IBM Backs Apache Spark For Big Data Analytics

Paul Miller, CONTRIBUTOR

FOLLOW ON FORBES (18)

"[IBM's initiatives] include:

- deepening the integration between Apache Spark and existing IBM products like the Watson Health Cloud;
- open sourcing IBM's existing SystemML machine learning technology;

http://aptuz.com/blog/is-apache-spark-going-to-replace-hadoop/

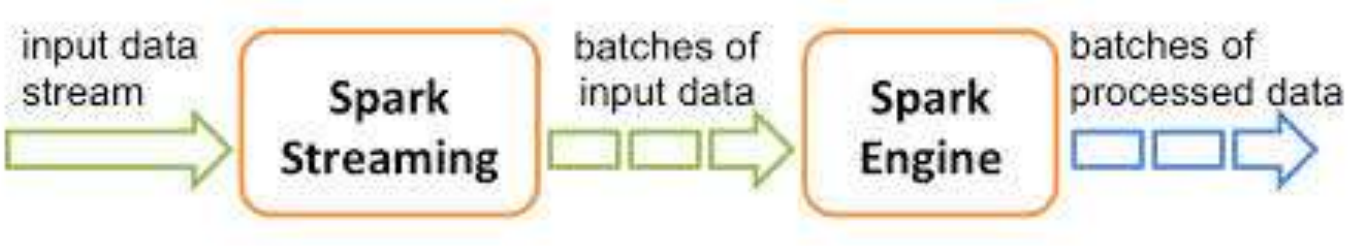
http://fortune.com/2015/09/09/cloudera-spark-mapreduce/

http://www.ebaytechblog.com/2014/05/28/using-spark-to-ignite-data-analytics/

http://www.forbes.com/sites/paulmiller/2015/06/15/ibm-backs-apache-spark-for-big-data-analytics/

Spark Streaming

- is no real streaming solution
- uses micro-batches
- cannot process data in real-time (i.e. no ultra-low latency)
- allows easy combination with other Spark components (SQL, Machine Learning, etc.)



Apache Spark – Hello World

Word Count

In this example, we use a few more transformations to build a dataset of (String, Int) pairs called `counts` and then save it to a file.

Python | Scala | Java

```
val textFile = spark.textFile("hdfs://...")
val counts = textFile.flatMap(line => line.split(" "))
                 .map(word => (word, 1))
                 .reduceByKey(_ + _)
counts.saveAsTextFile("hdfs://...")
```

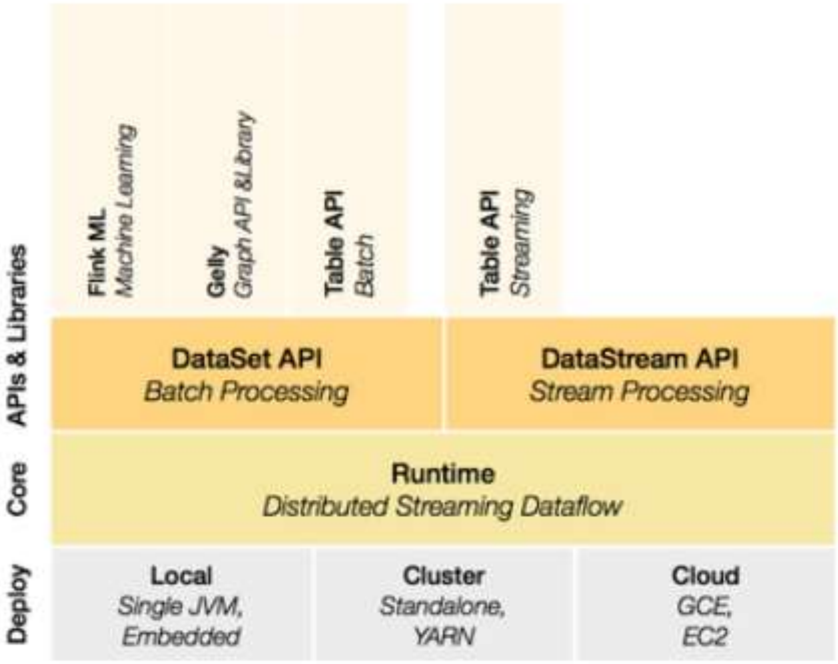
← Spark Core API

Spark Streaming API →

```
TwitterUtils.createStream(...)
    .filter(_.getText.contains("Spark"))
    .countByWindow(Seconds(5))
```

Counting tweets on a sliding window

TIBCO | Apache Flink



Spark Streaming

- „Newcomer“
- Looks very similar to Spark
- But „Streaming First“ concept

Streaming First

High throughput and low latency stream processing with exactly-once guarantees.

Batch on Streaming

Batch processing applications run efficiently as special cases of stream processing applications.

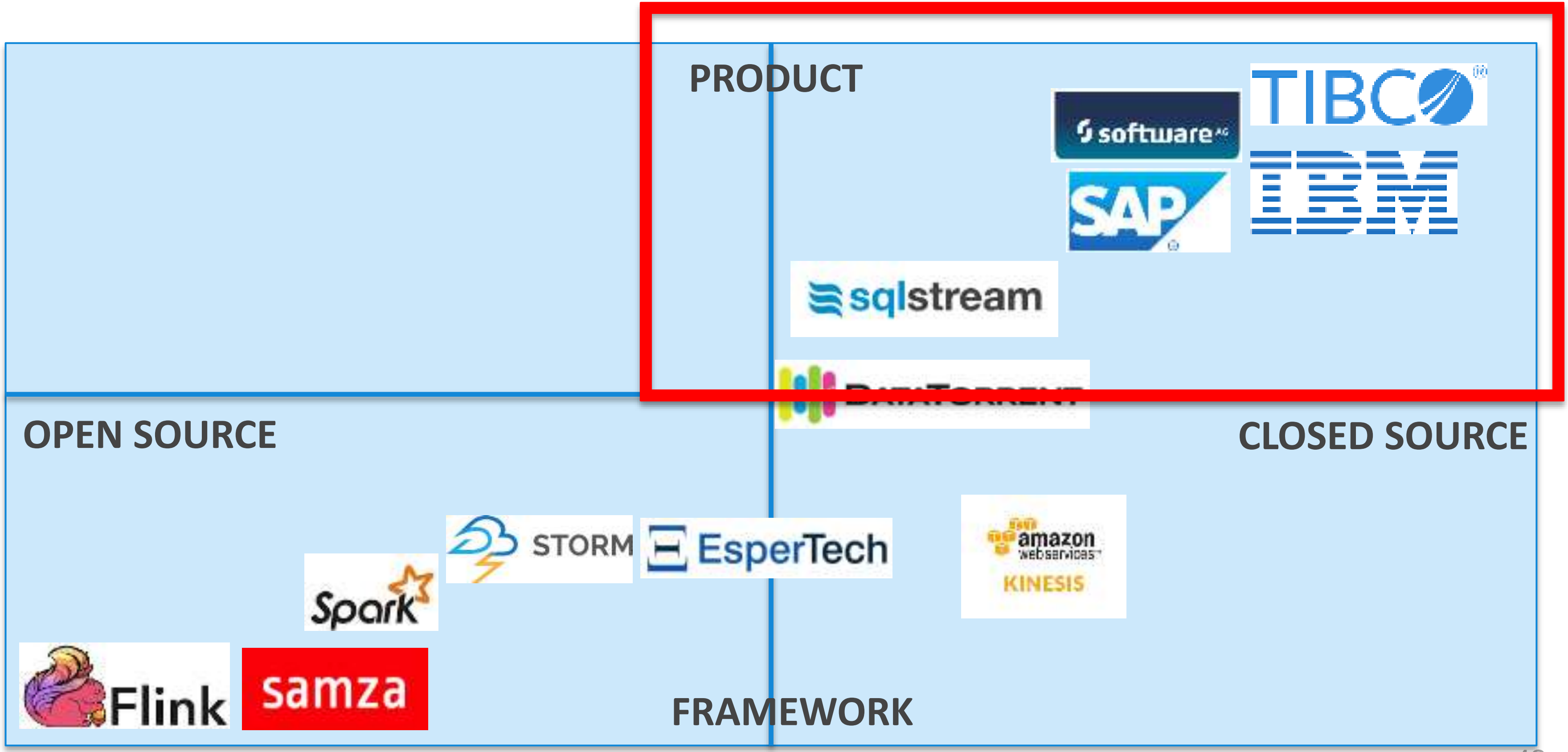
TIBCO | Apache Flink vs. MapReduce, Tez, Spark

hadoop MapReduce	TEZ	Spark	Apache Flink
• Batch	• Batch • Interactive	• Batch • Interactive • Near-Real time Streaming • Iterative processing	• Batch • Interactive • **Real-Time Streaming** • **Native Iterative** processing
MapReduce	**D**irect **A**cyclic **G**raphs (DAG) Dataflows	RDD: **R**esilient **D**istributed **D**atasets	**Cyclic Dataflows**
1st Generation (1G)	2nd Generation (2G)	3rd Generation (3G)	4th Generation (**4G**)

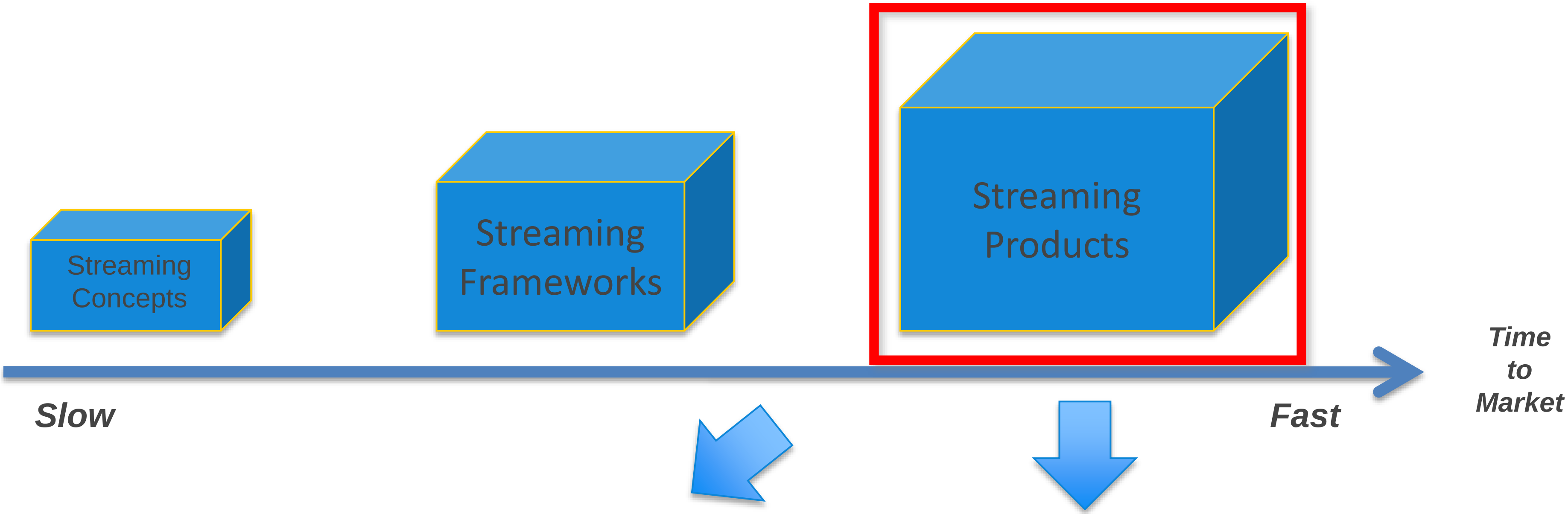
http://www.slideshare.net/sbaltagi/overview-of-apacheflinkbyslimbaltagi/12

Alternatives

(no complete list!)



What Streaming Alternative do you need?



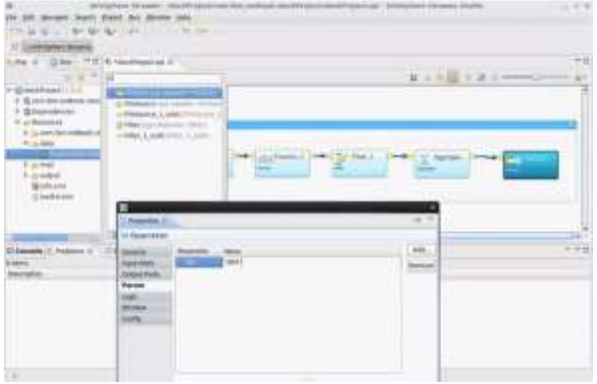
Library (Java, .NET, Python)
Query Language (often similar to SQL)
Scalability (horizontal and vertical, fail over)
Connectivity (technologies, markets, products)
Operators (Filter, Sort, Aggregate)

Visual IDE (Dev, Test, Debug)
Simulation (Feed Testing, Test Generation)
Live UI (monitoring, proactive interaction)
Maturity (24h support, consulting)
Integration (out-of-the-box: ESB, MDM, etc.)

IBM InfoSphere Streams: An open platform

Highlights

- Adopt a shareable, open and accessible platform that is widely used for stream processing
- Facilitate developer productivity and enable deep analytics for the business to accelerate time to market and maximize value
- Capitalize on an open platform to simplify operations and streamline integration with existing data management tools
- Reduce risks with a solution supported by IBM and thriving communities
- Maximize flexibility with Quick Start, Developer Edition, Production Edition and Cloud offerings plus a variety of pricing models



Of Streams and Storms

A Direct Comparison of IBM InfoSphere Streams and Apache Storm in a Real World Use Case – Email Processing

Zubair Nabi and Eric Bouillet
IBM Research Dublin

Andrew Bainbridge and Chris Thomas
IBM Software Group Europe

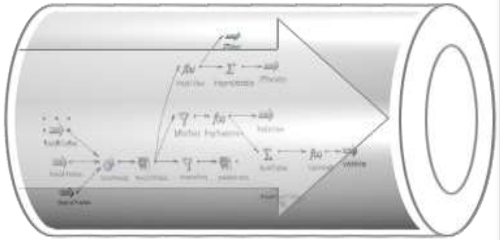
April 2014

In this paper, we compare the performance of IBM InfoSphere Streams against Apache Storm [1], a leading open source alternative, to augment existing literature [2]. To this end, we implemented a real-world stream processing application, which enables email classification for online spam detection [3] on both platforms. Our goal was to analyze both the quantitative differences in performance as well as the qualitative differences in application writing and framework tuning. Similar to other studies [4, 5], we employed CPU time and throughput as primary metrics to compare the efficacy of both systems. Overall, our results show that for the application benchmark documented in this paper, Streams outperforms Storm by **2.6 to 12.3** times in terms of throughput while simultaneously consuming **5.5** to **14.2** times less CPU time.

https://developer.ibm.com/streamsdev/wp-content/uploads/sites/15/2014/04/Streams-and-Storm-April-2014-Final.pdf

TIBCO StreamBase

- **Performance: Latency, Throughput, Scalability**
  - Multi-threaded and clustered server from version 1
  - High throughput: Millions of messages, 100,000s of quotes, 10,000s of orders
  - Low-latency: microsecond latency for algo trading, pre-trade risk, market data
- **Take Advantage of High Performance Hardware**
  - Multicore (12, 24, 32 core) large memory (10s of gigabytes)
  - 64-bit Linux, Windows, Solaris deployment
  - Hardware acceleration (GPU, Solace, Tervela)
- **Enterprise Deployment**
  - High availability and fault tolerance
  - Distributed state management for large data sets
  - Management and monitoring tools
  - Security and entitlements Integration
  - Continuous deployment and QA Process Support

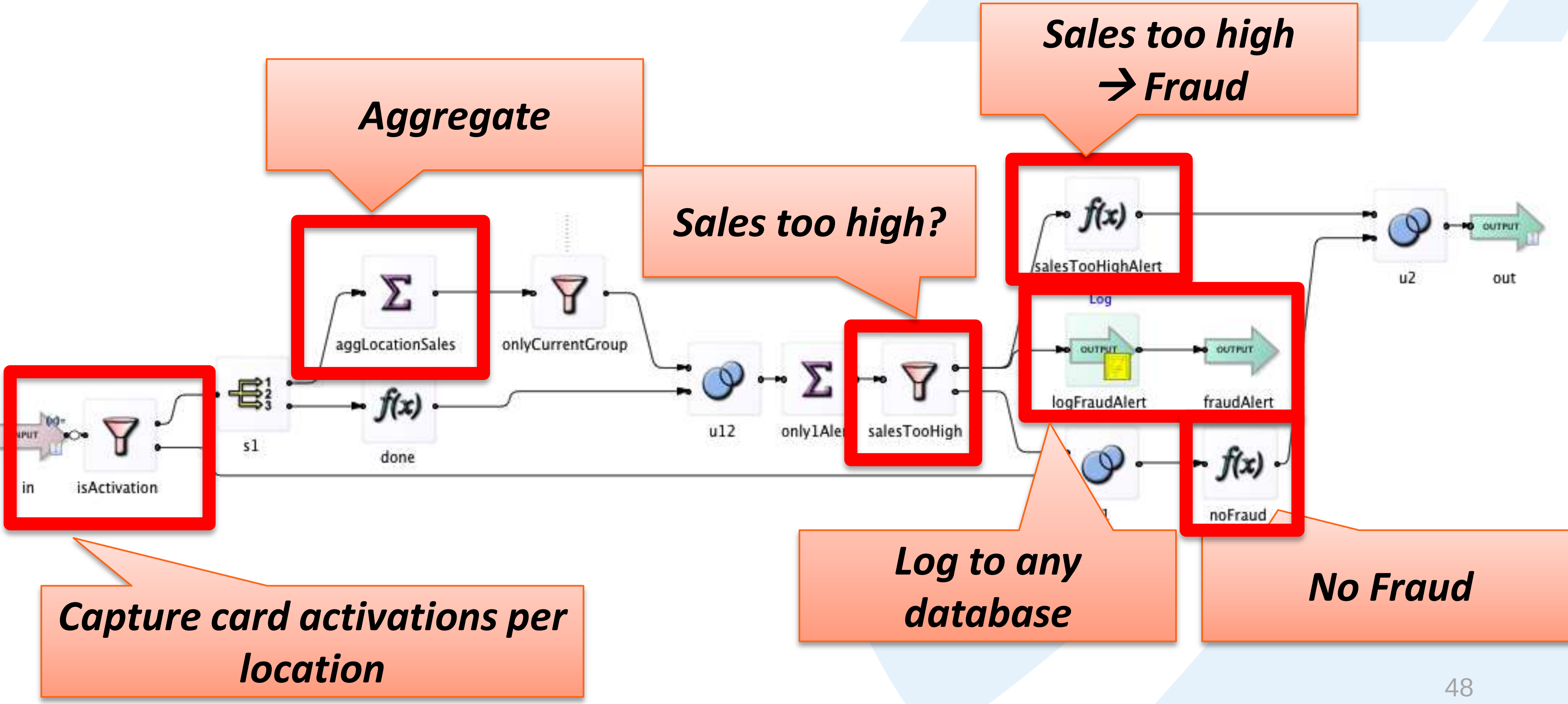


"The StreamBase engine is for real. We couldn't break it, and believe me, I tried" SVP Development, Top 5 Broker Dealer

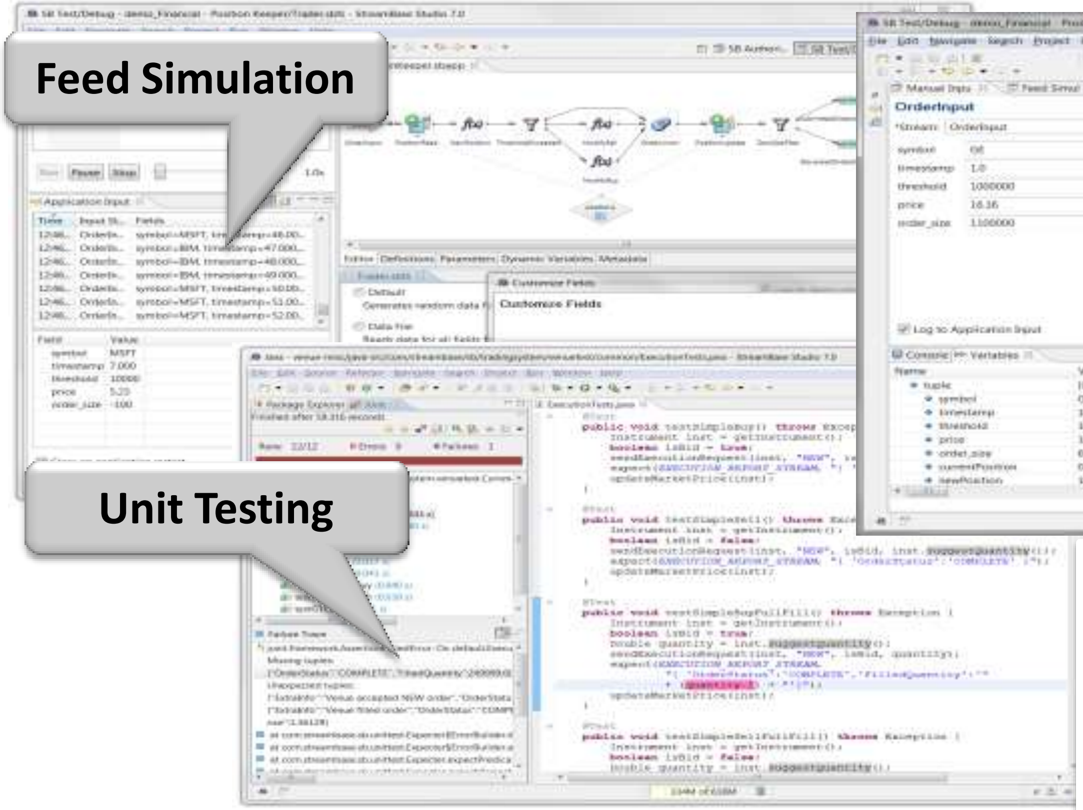
StreamBase Server Innovations

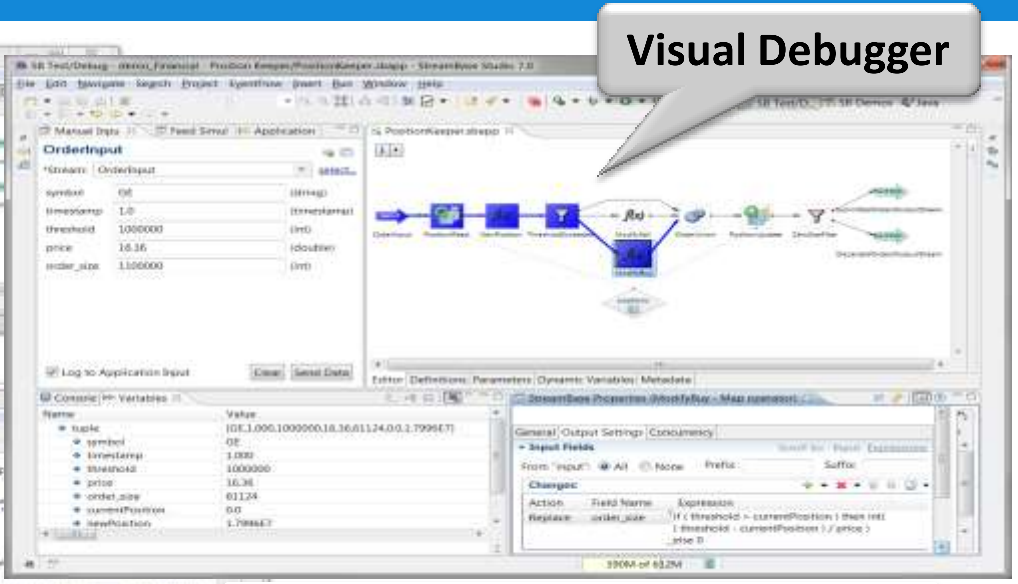
- StreamSQL compiler and static optimizer
- In process, in thread adapter architecture
- Visual parallelism and scaling
- Data parallelism and dispatch
- ActiveSpaces integration for distributed shared state

StreamBase: The Power of Visual Programming



StreamBase Development Studio



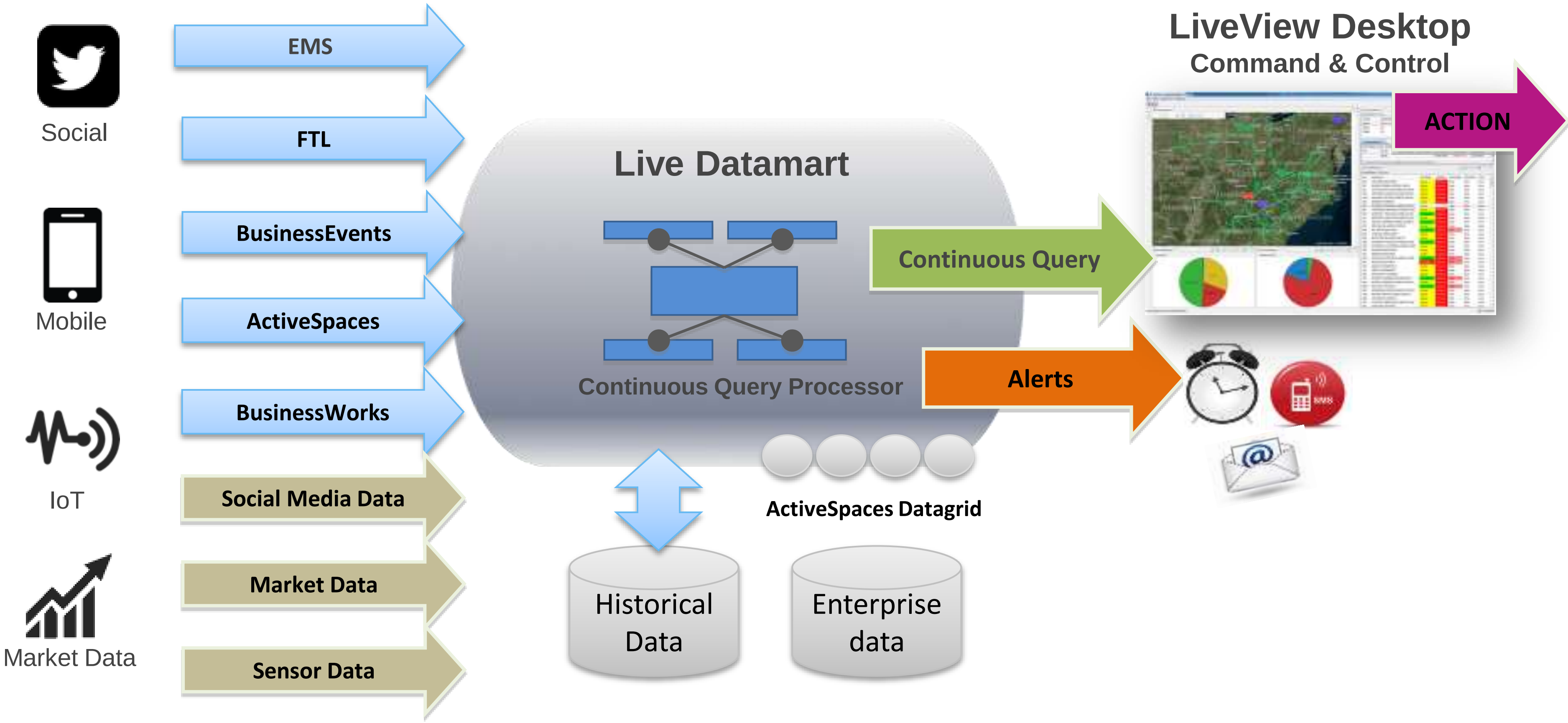


"StreamBase's modeling tools are easy to use and will enable the exchange to quickly react to the ever changing needs of our customers." *Steve Goldman, Director of Enterprise Architecture*

CME Group

A CME/Chicago Board of Trade/NYMEX Company

TIBCO | Live Datamart



TIBCO | LiveView Desktop

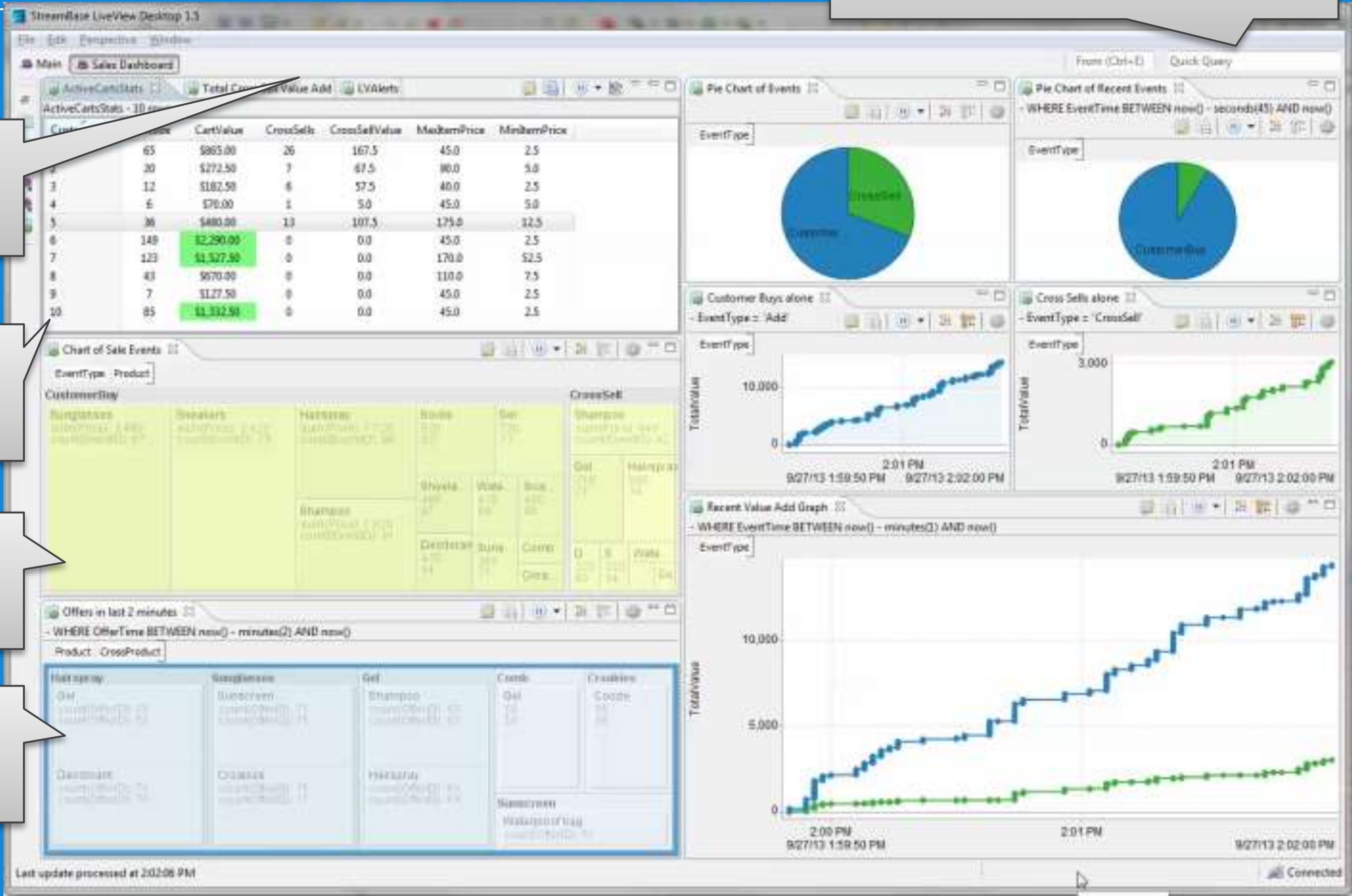
Ad-hoc continuous query

Alerts

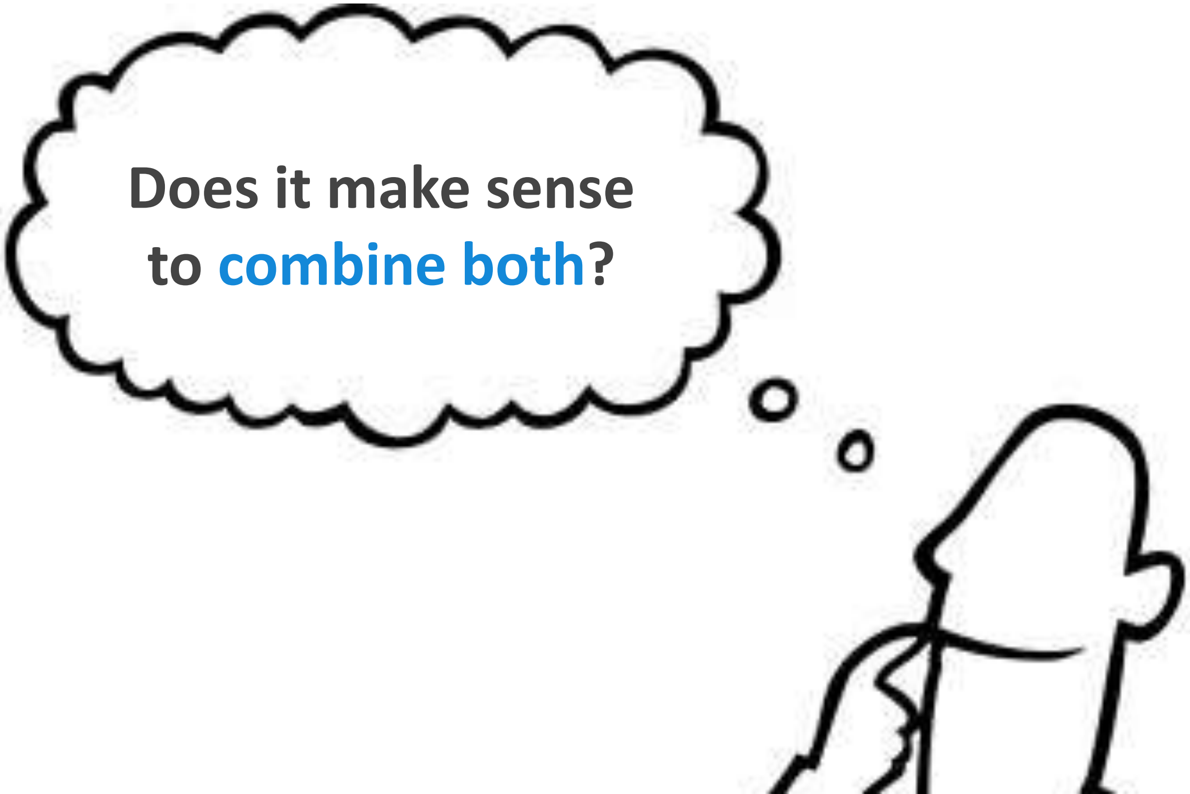
Dynamic aggregation

Action

Live visualization

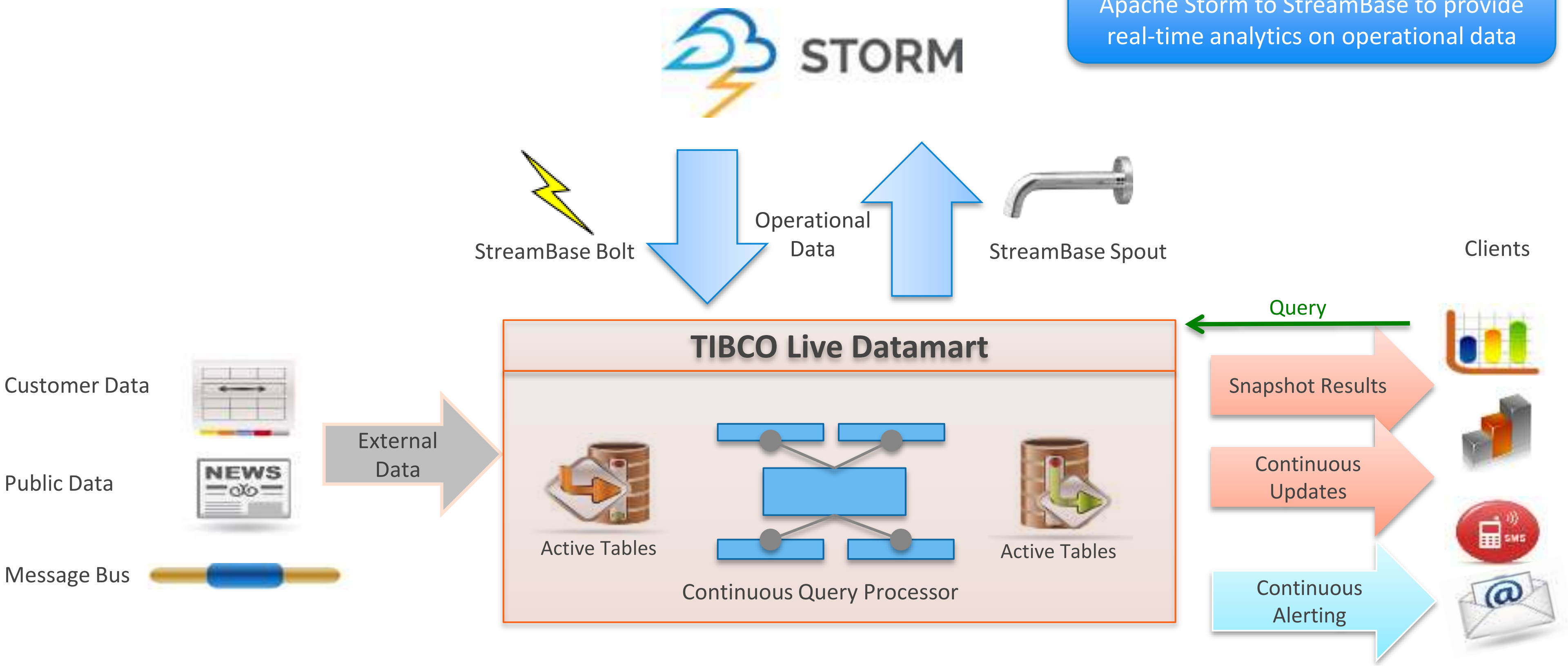


Spoilt for Choice – Framework or Product?



Example: Apache Storm + TIBCO Live Datamart

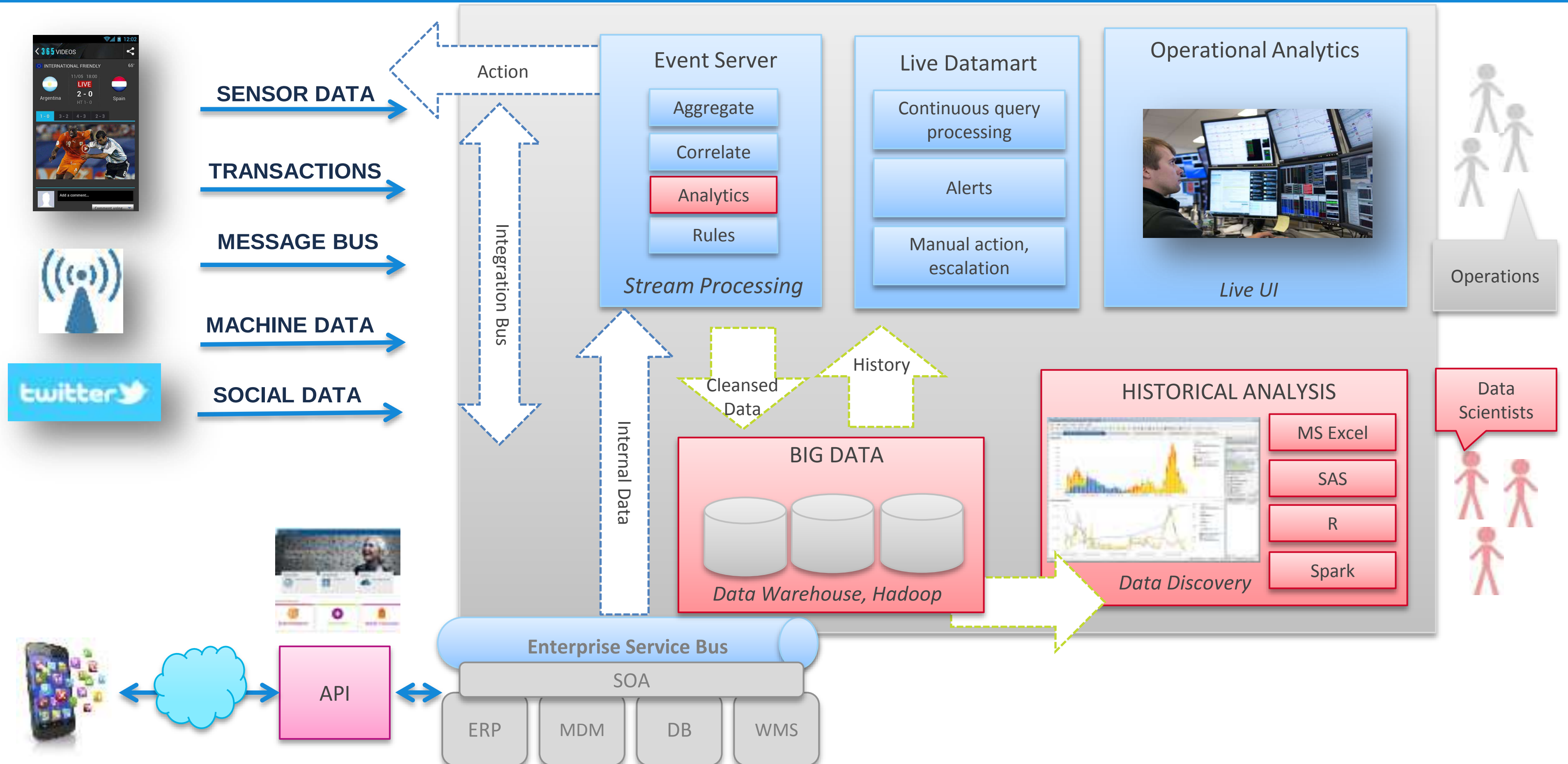
StreamBase Bolt and Spout connect Apache Storm to StreamBase to provide real-time analytics on operational data



TIBCO™ |
Agenda

– Real World Use Cases
– Introduction to Stream Processing
– Market Overview
– **Relation to other Big Data Components**
– Live Demo

TIBCO | Streaming Analytics Reference Architecture



Real Time Close Loop

Analyze

Analyze data via Data Discovery

Uncover patterns, trends, correlations

Model

Develop model

Deploy into Stream Processing flow

Act

Automatically monitor real-time transactions

Automatically trigger action

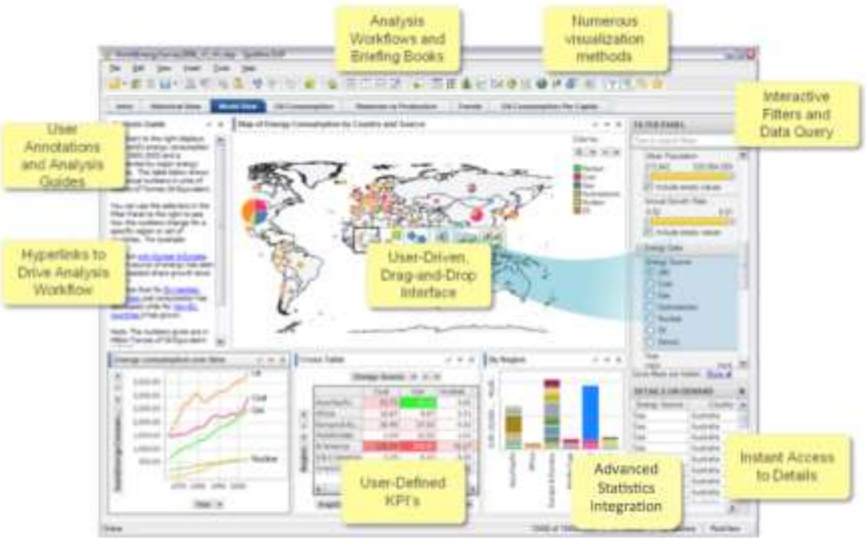
Big Data

- store everything
  in Hadoop, DWH, NoSQL, etc.
- even without structure
- even if you do not need it today

http://blogs.teradata.com/international/tag/hadoop/

Data Discovery + Statistics + Machine Learning

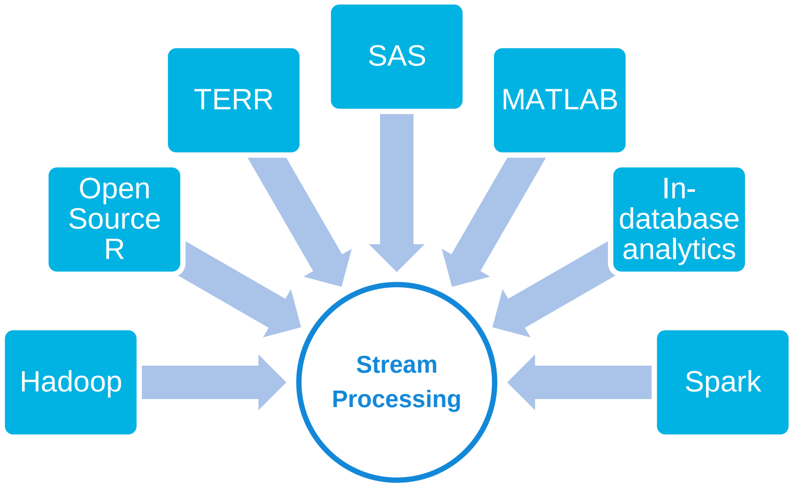
to find insights and patterns in historical data



Streaming Analytics

to operationalize insights

and patterns in real time



TODAY

80% of betting happens

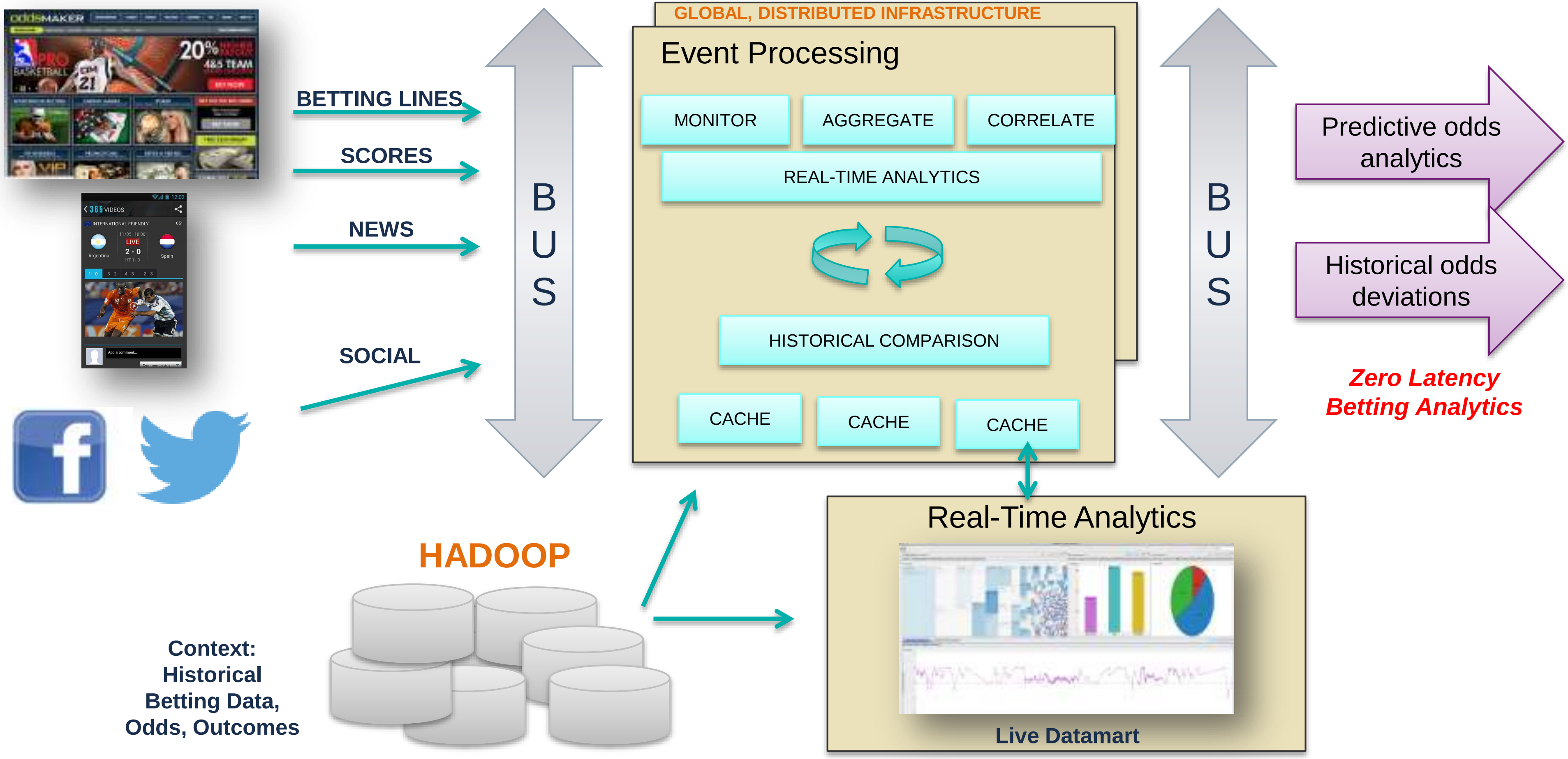
AFTER the game begins

TIBCO | Case Study: Streaming Analytics for Betting

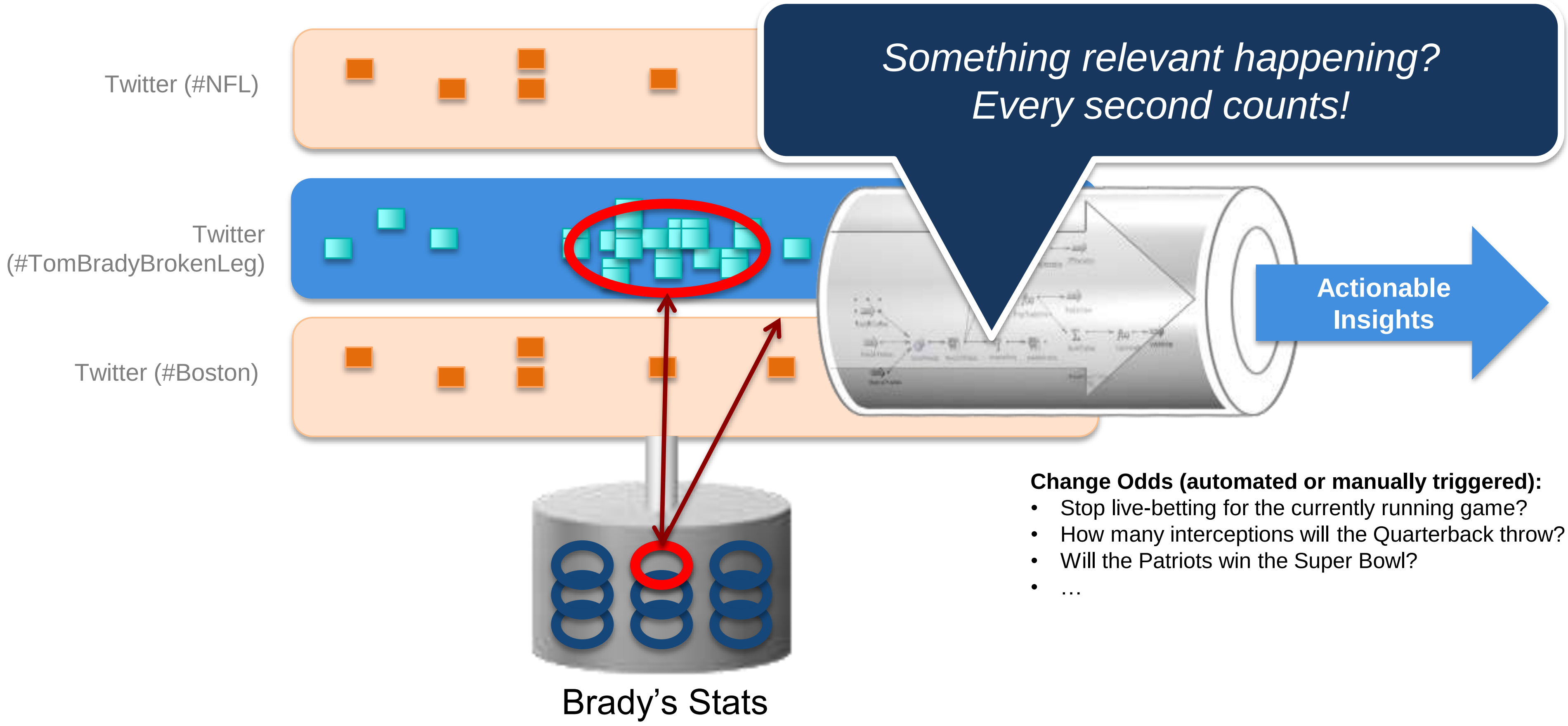
"With StreamBase, in two months we had our first betting analytics feed live, and we continually deploy new ideas and evolve our old ones."
- Alex Kozlenkov, VP of technology, TXOdds

- **Situation: Today, 80% of Betting is Done *After* the Game Starts**
  - It's not your father's bookie anymore!
- **Problem: How to Analyze Big Betting Data?**
  - Thousands of concurrent games, constantly adjusting odds, dozens of betting networks – firms must correlate millions of events a day to find the best betting opportunities in real-time
- **Solution: TIBCO for Fast Data Architecture**
  - TXOdds uses TIBCO to correlate, aggregate, and analyze large volumes of streaming betting data in real-time and publish innovative predictive betting analytics to their customers
- **Result: TXOdds First to Market with Innovative Zero Latency Betting Analytics**
  - Innovative real-time analytics help players who can process electronic data in real-time the edge

TIBCO™ | Reference Architecture: Streaming Betting Analytics



TIBCO | Real-Time Social Media Analytics



Change Odds (automated or manually triggered):

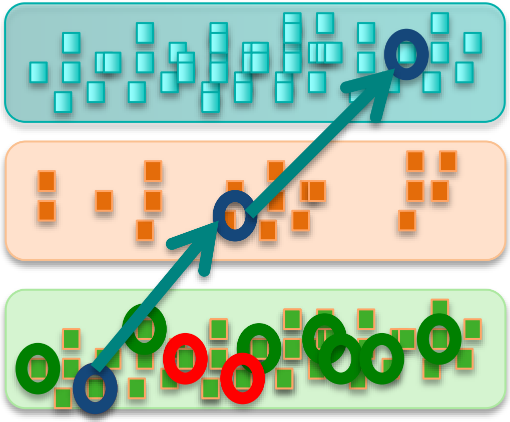
- Stop live-betting for the currently running game?
- How many interceptions will the Quarterback throw?
- Will the Patriots win the Super Bowl?
- …

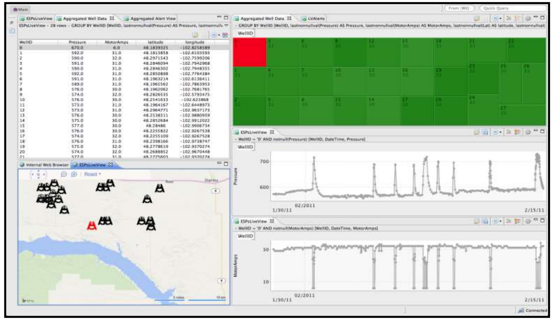
Real-Time Social Sentiment Analysis

Agenda

- Real World Use Cases
- Introduction to Stream Processing
- Market Overview
- Relation to other Big Data Components
- **Live Demo**

TIBCO® | Live Demo





Predictive Sensor Analytics

Did you get the Key Message?

Key Messages

- Streaming Analytics processes Data while it is in Motion!
- Automation and Proactive Human Interaction are BOTH needed!
- Time to Market is the Key Requirement for most Use Cases!

Questions?

Kai Wähner

kwaehner@tibco.com

@KaiWaehner

www.kai-waehner.de

LinkedIn / Xing → Please connect!