



# **BUILDING BIG DATA OPERATIONAL INTELLIGENCE PLATFORM WITH APACHE SPARK**

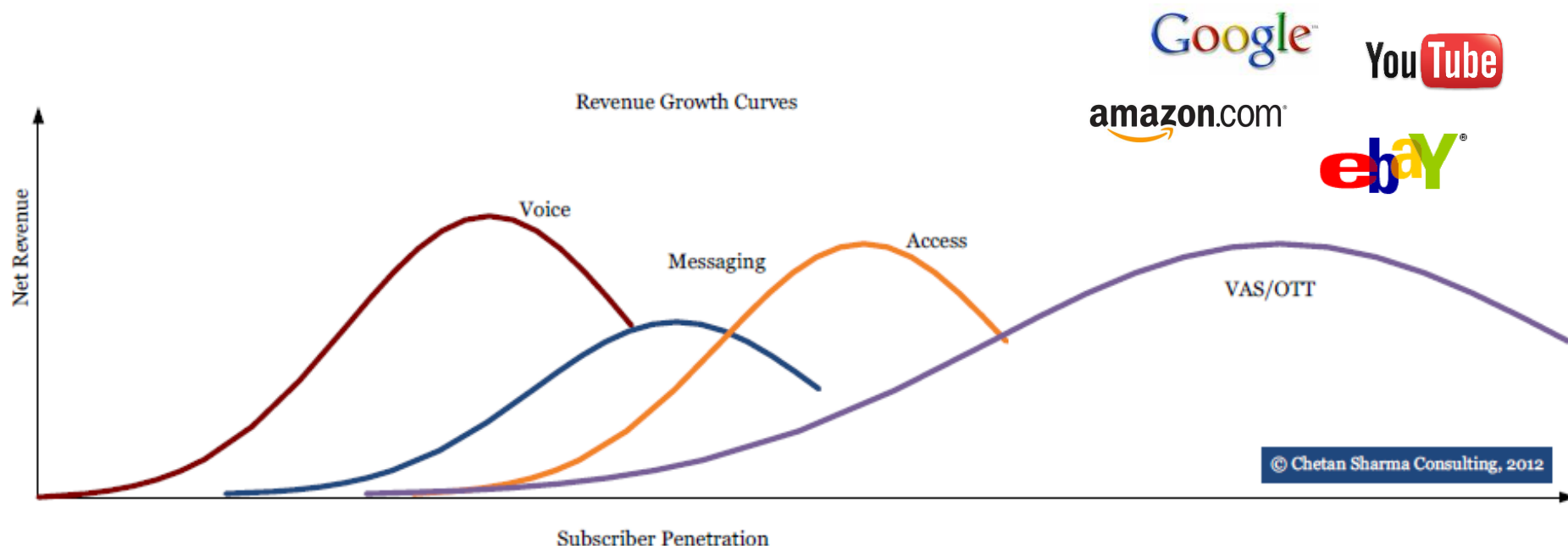
Eric Carr (VP Core Systems Group)  
Spark Summit 2014

# Communication Service Providers & Big Data Analytics

---

*Market & Technology Imperatives*

# Industry Context for Communication Service Providers (CSPs)



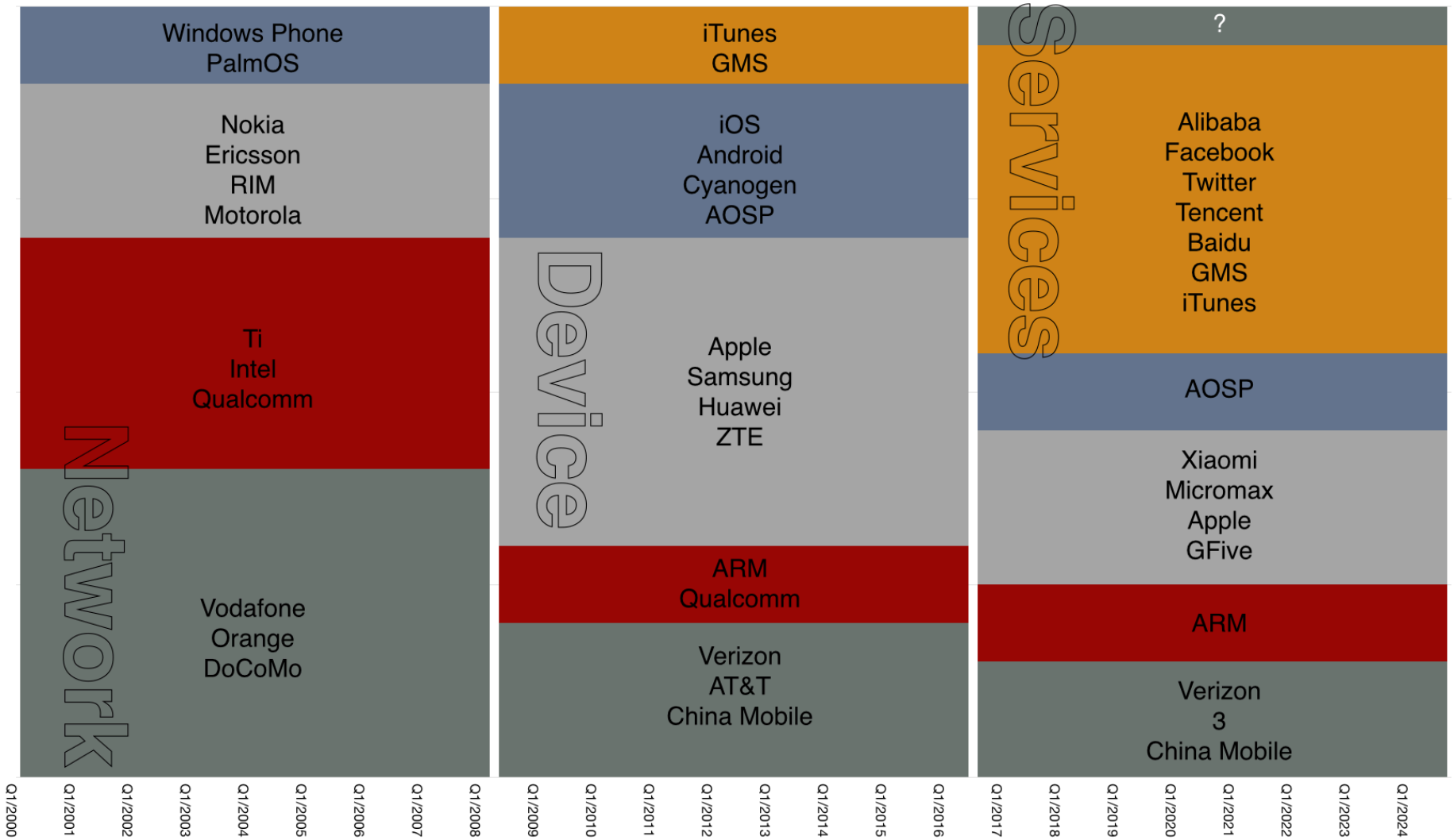
## Big Data is at the heart of two core strategies for CSPs:

- Improve current revenue sources through greater operational efficiencies
- Create new revenue source with the 4th wave

Source: Chetan Consulting

# CSPs - Industry Value Chain Shift

Distribution of Profit by Era



# CSPs - A High Bar for Operational Intelligence



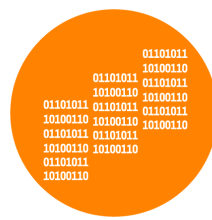
## Exponential Data Growth

Petabytes of data per day;  
billions of records per day



## Distributed Network

Dozens of locations for capturing data, scattered around a vast territory



## Data Diversity

Hundreds of sources from different equipment types and vendors



## Timely Insights

Automated reactions triggered in seconds

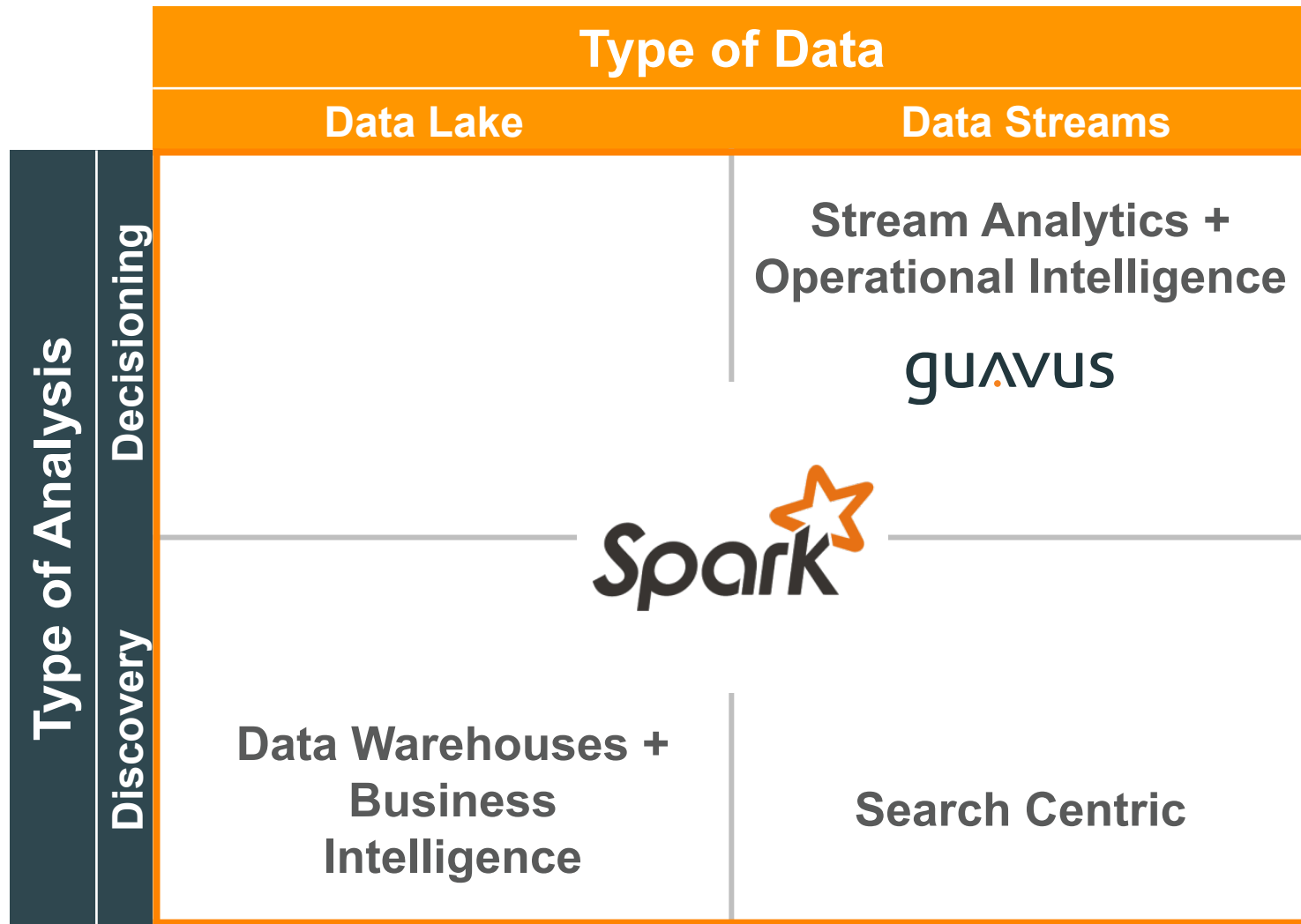


## High Availability

No data loss;  
no down time

**CSPs require solutions engineered to meet very stringent requirements**

# Different Platforms target Different Questions



# Streaming Analytics & Machine Learning to Action

---

# Driving Streaming Analytics to Action

**Network Flow Analytics**



**Usage Awareness**



**Operational Interactions**



**Real-Time Actions**



NetFlow, Routing Planning



Layer 7 Visibility



Operational Intelligence



Content & CDN Analytics



Policy Profile Triggers



Care & Experience Mgmt



New Service Creation & Monetization



Small Cell / RAN / Backhaul Differentiation



SON / SDN / Virtualization



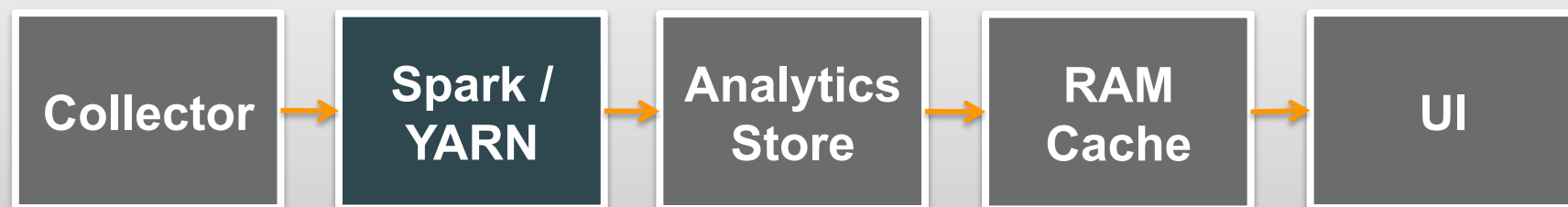
Content Optimization



## Reflex 1.0 Pipeline – Timely Cube Reporting



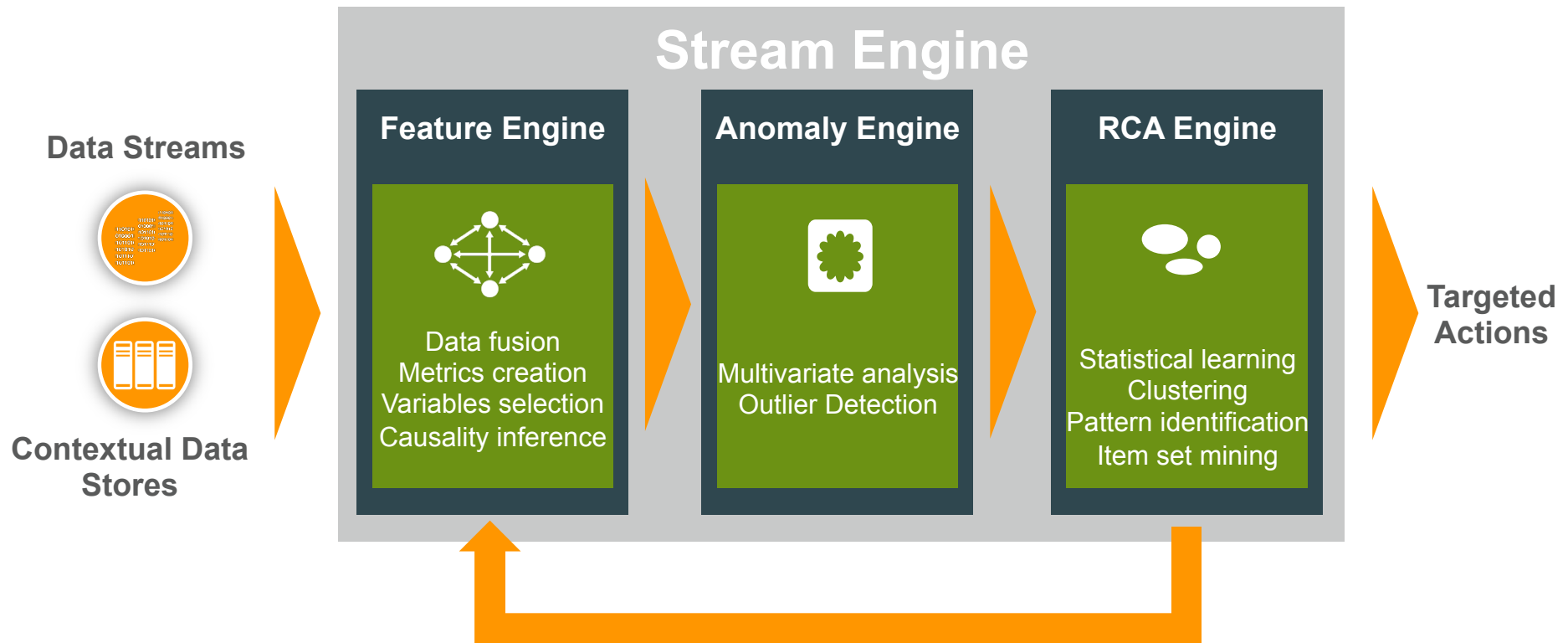
## Reflex 1.5 Pipeline – Spark / Yarn



## Reflex 2.0 Pipeline – Spark Streaming core



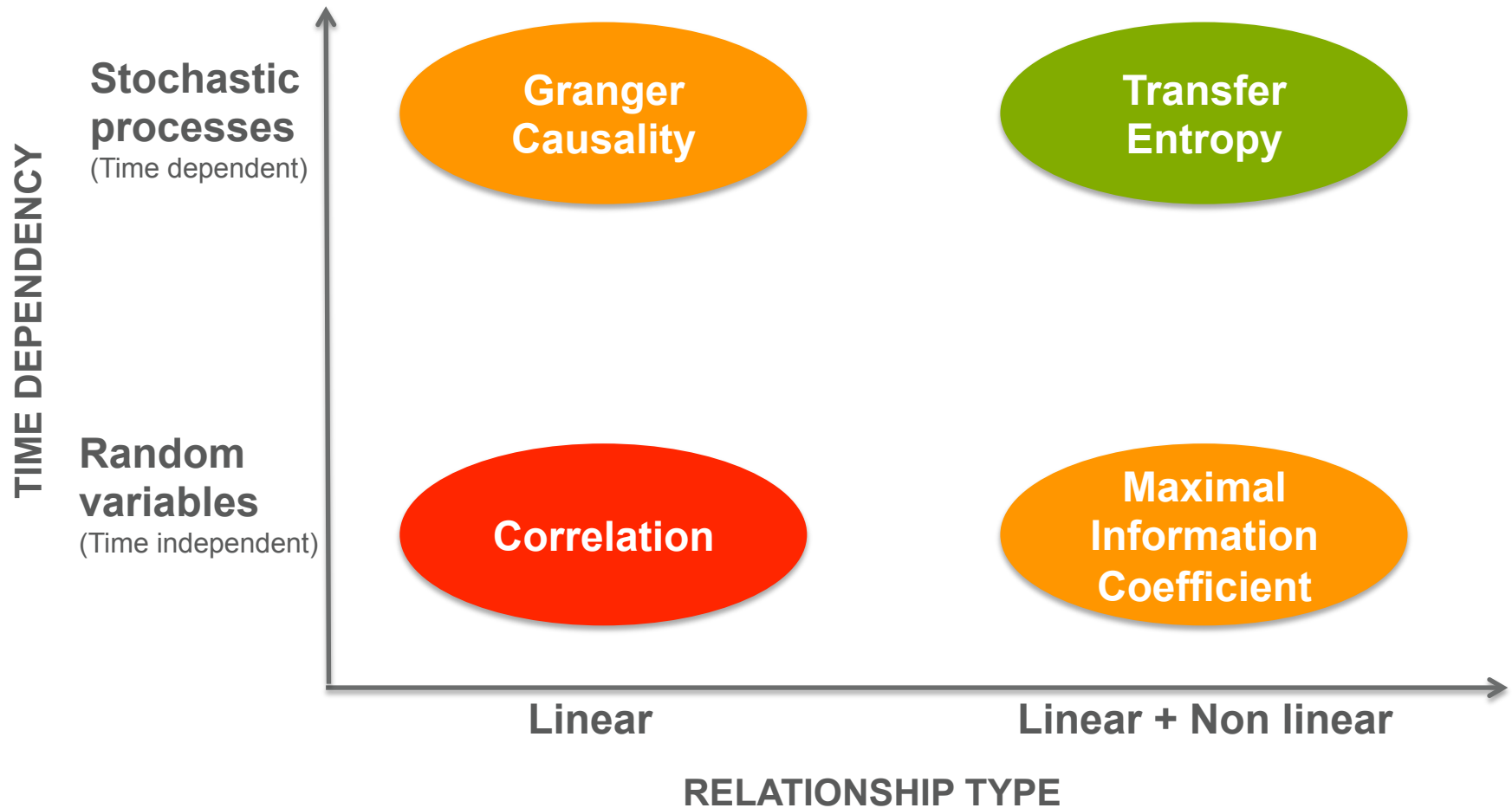
# Stream Engine - Operational Intelligence Analytics



## Optimized algorithmic support for common stream data processing & fusions

- Detect unusual events occurring in stream(s) of data. Once detected, isolate root cause
- Anomaly / outlier detection, commonalities / root cause forensics, prediction / forecasting, actions / alerts / notifications
- Record enrichment capabilities – e.g. URL categorization, device id, etc.

# Example - State of the art causality analysis



**Ranking of metrics:**

- 1) Transfer Entropy
- 2) Maximal Information Coefficient, Granger Causality
- 3) Correlation

# Example - Causality Techniques

**Transfer entropy** from a process  $X$  to another process  $Y$  is the amount of uncertainty reduced in future values of  $Y$  by knowing the past values of  $X$  given past values of  $Y$ .

## PROS

Model free, information theory based approach

Most generic estimation of causality between two random processes

## CONS

Challenging joint probability estimation

Large amount of data needed for calculation

Choice of time lags

**Maximal Information Coefficient** is a measure of the strength between two random variables based on mutual information. Methodology for empirical estimation based on maximizing the mutual information over a set of grids

## PROS

Model free, information theory based approach

Can find linear and non-linear relationships

Estimation possible with smaller dataset

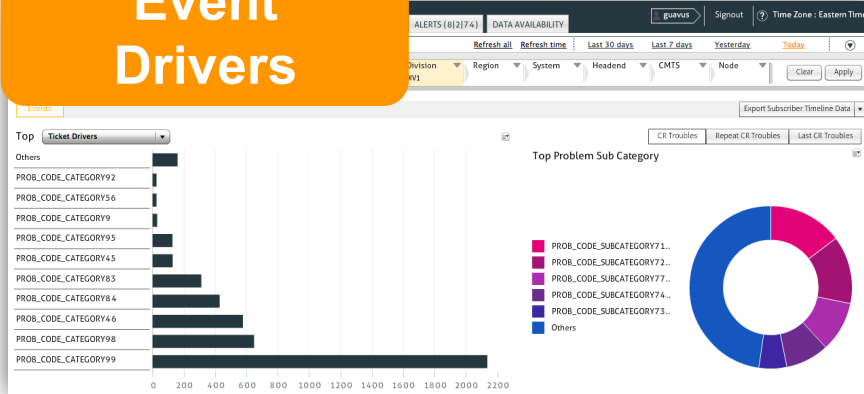
## CONS

No time information

# Network Operations / Care Example

## Identifying Commonalities, Anomalies, RCA

### Event Drivers



### Anomaly Detection

Alert Name	Time	Frequency	Severity	Measure	Limit Type	Value	Threshold	Action
Truck rolls - Presidio	14-Mar 15:00	Hourly	High	Truck rolls	Week on Week % Deviation	(65)18.5%	90%	<a href="#">Dismiss</a> <a href="#">Dismiss All</a> <a href="#">Commonalities</a>
Non ticketed calls - San Jacinto	14-Mar 15:00	Hourly	Low	Non-Ticketed Calls	Absolute	846	400	<a href="#">Dismiss</a> <a href="#">Dismiss All</a> <a href="#">Commonalities</a>
Calls - Agent Handled - Desoto	14-Mar 15:00	Hourly	High	Non-Ticketed Calls	30 Day Baseline % Deviation	(59)136%	90%	<a href="#">Dismiss</a> <a href="#">Dismiss All</a> <a href="#">Commonalities</a>
Video Issue tickets - Jefferson	14-Mar 14:00	Hourly	Medium	Ticketed Calls	Week on Week % Deviation	(61)79%	50%	<a href="#">Dismiss</a> <a href="#">Dismiss All</a> <a href="#">Commonalities</a>
Calls - Agent Handled - Sarasota	14-Mar 14:00	Hourly	Low	Non-Ticketed Calls	Absolute	998	400	<a href="#">Dismiss</a> <a href="#">Dismiss All</a> <a href="#">Commonalities</a>
Truck rolls - Wakulla	14-Mar 14:00	Hourly	High	Truck rolls	30 Day Baseline % Deviation	(155)158%	90%	<a href="#">Dismiss</a> <a href="#">Dismiss All</a> <a href="#">Commonalities</a>
DR400 - New issues - Calls	14-Mar 13:00	Hourly	Low	Non-Ticketed Calls	Absolute	1088	400	<a href="#">Dismiss</a> <a href="#">Dismiss All</a> <a href="#">Commonalities</a>
Platform rollout - new issues - Orange	14-Mar 13:00	Hourly	High	Ticketed Calls	Week on Week % Deviation	(120)103%	90%	<a href="#">Dismiss</a> <a href="#">Dismiss All</a> <a href="#">Commonalities</a>
Calls over limit - Calhoun	14-Mar 13:00	Hourly	High	Non-Ticketed Calls	30 Day Baseline % Deviation	(324)181%	90%	<a href="#">Dismiss</a> <a href="#">Dismiss All</a> <a href="#">Commonalities</a>
Calls over limit - Galveston	14-Mar 12:00	Hourly	Low	Non-Ticketed Calls	Absolute	1113	400	<a href="#">Dismiss</a> <a href="#">Dismiss All</a> <a href="#">Commonalities</a>
Calls - Agent Handled - Sarasota	14-Mar 12:00	Hourly	High	Non-Ticketed Calls	30 Day Baseline % Deviation	(646)141%	90%	<a href="#">Dismiss</a> <a href="#">Dismiss All</a> <a href="#">Commonalities</a>
Tickets - VOIP - Marta Gorda	14-Mar 11:00	Hourly	Low	Non-Ticketed Calls	Absolute	1084	400	<a href="#">Dismiss</a> <a href="#">Dismiss All</a> <a href="#">Commonalities</a>
DR400 - New issues - Calls	14-Mar 11:00	Hourly	Low	Ticketed Calls	Week on Week % Deviation	(259)32%	20%	<a href="#">Dismiss</a> <a href="#">Dismiss All</a> <a href="#">Commonalities</a>
Calls - Agent Handled - Desoto	14-Mar 10:00	Hourly	Low	Non-Ticketed Calls	30 Day Baseline % Deviation	(955)51%	25%	<a href="#">Dismiss</a> <a href="#">Dismiss All</a> <a href="#">Commonalities</a>
Test Alert - No action	14-Mar 10:00	Hourly	Low	Non-Ticketed Calls	Absolute	1019	400	<a href="#">Dismiss</a> <a href="#">Dismiss All</a> <a href="#">Commonalities</a>
Calls - Agent Handled - St.Johns	14-Mar 09:00	Hourly	Low	Non-Ticketed Calls	Absolute	1047	400	<a href="#">Dismiss</a> <a href="#">Dismiss All</a> <a href="#">Commonalities</a>
Video Issue tickets - Jefferson	14-Mar 09:00	Hourly	Low	Ticketed Calls	Week on Week % Deviation	(299)21%	20%	<a href="#">Dismiss</a> <a href="#">Dismiss All</a> <a href="#">Commonalities</a>

### Event Chaining



### Root-Cause Analysis

Customer Reported Tickets	Top Commonalities	Relative
CR Solution Code	<input checked="" type="checkbox"/> A 311:30L:CODE_SOLDDESCR91	
CR Solution Code Subcategory	<input checked="" type="checkbox"/> B SOL:CODE_SUBCATEGORY37	
CR Problem Code Subcategory	<input checked="" type="checkbox"/> C PROB_CODE_SUBCATEGORY77A	
Set-Top Box Firmware	<input checked="" type="checkbox"/> D Firmware2	
Set-Top Box Model	<input checked="" type="checkbox"/> E OCT2000	

Time Range: 15 Mar 2013, 13:00 - 15 Mar 2013, 14:00

Top Combinations	Relative % of Case Events	Subscriber Impact
35.0% A	143	
35.0% B	143	
35.0% A, B	143	
13.5% C	55	
7.1% B, C	29	
7.1% A, B, C	29	
7.1% A, C	29	

# BinStream Details

---

# Use Case

- Use IP Traffic Records to calculate Bandwidth Usage as a Time Series (*continuously ...*), *can't do* that based on the time the records are received by Spark.
  - In general for any record which has a timestamp, it important to analyze based on the time of event rather than the reception of the event *record*.

# Challenges With

- For one dataset. Make the time stamp part of the key.



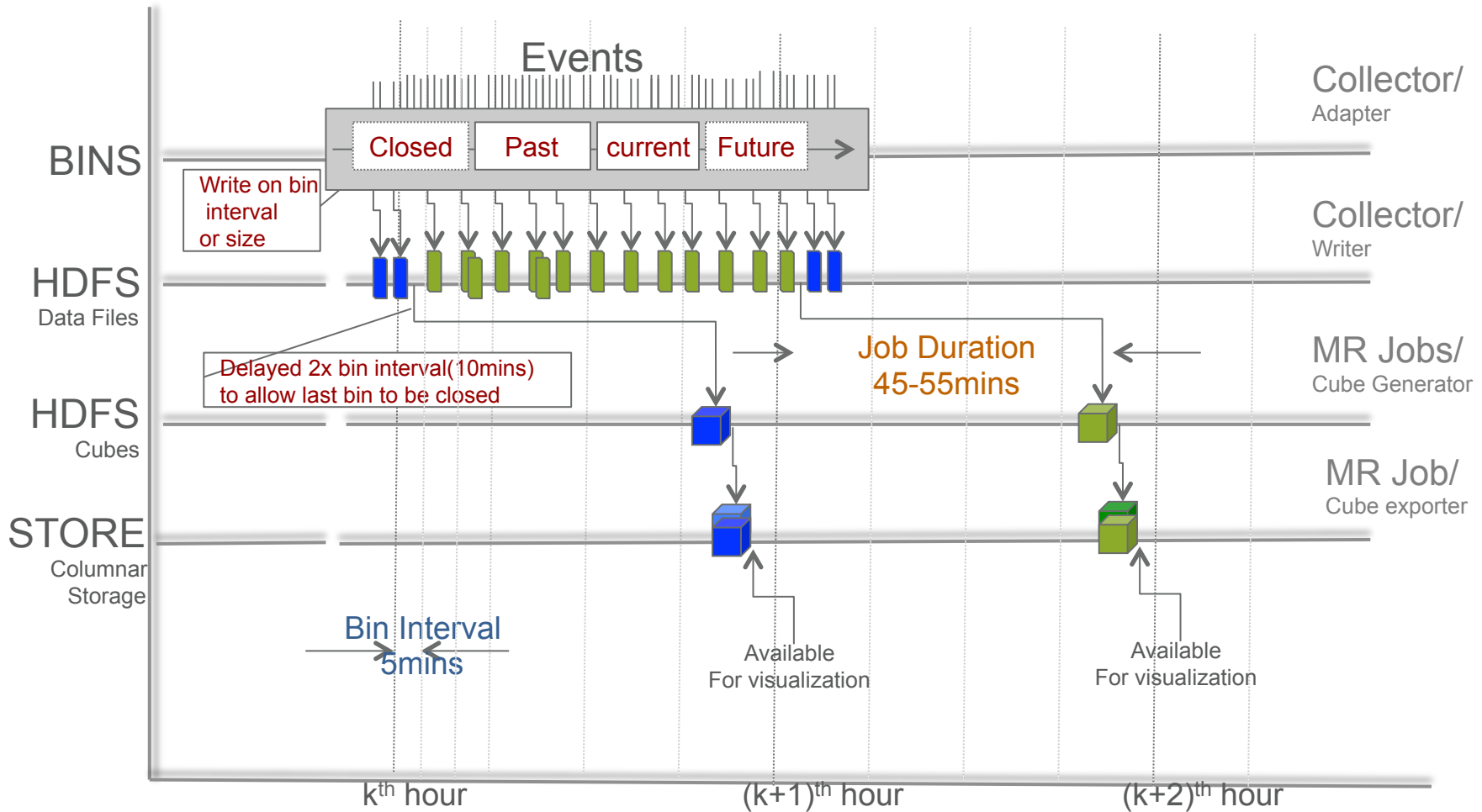
- For continuously streamed data sets.



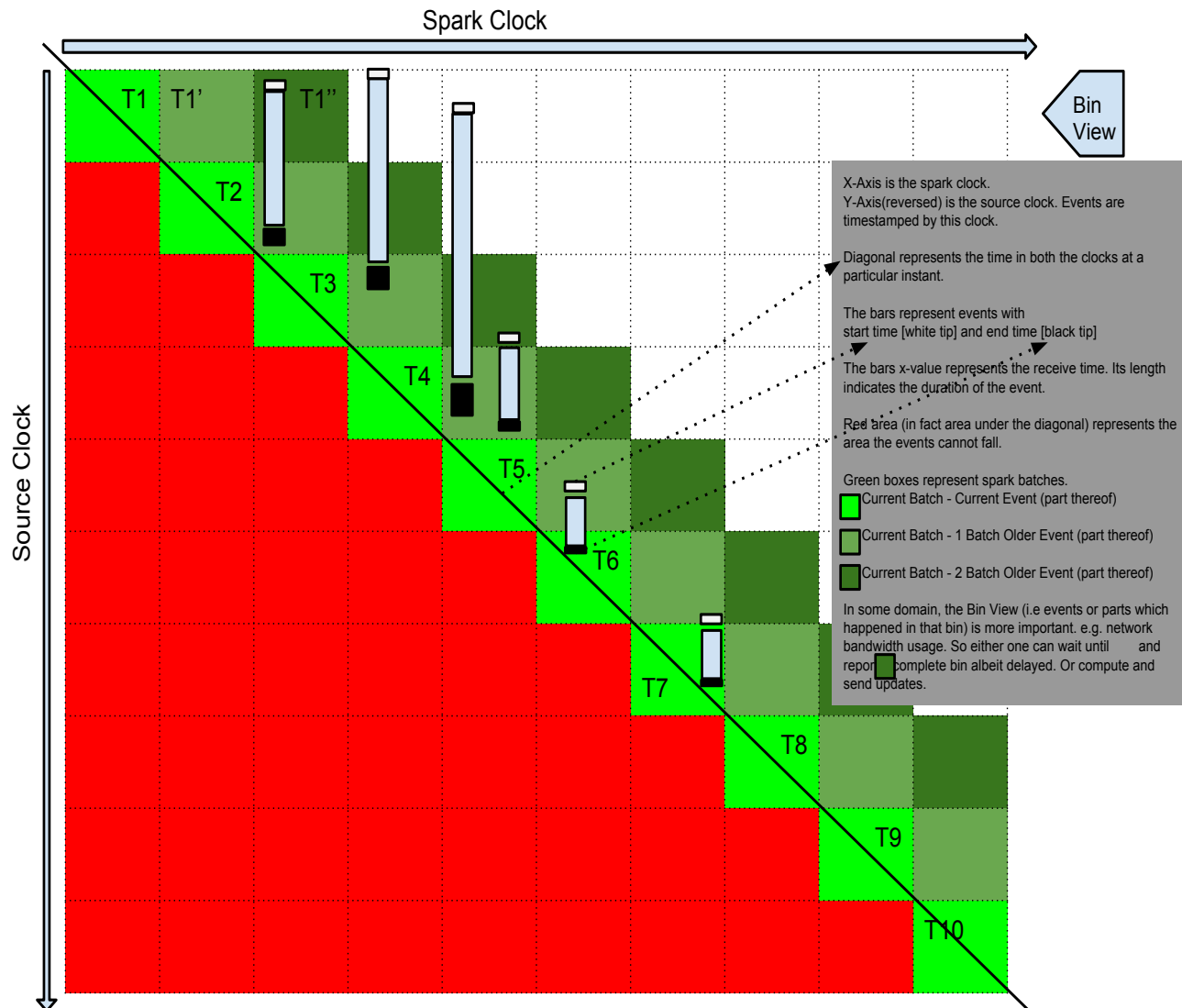
- You do not know if you have received all data for a particular time slot. Caused by event delay, or event duration.
- An event could span multiple time slots. Caused by event duration.



# Data Processing – Timing (Map-Reduce world)



# Proposed Binning & Proration Solution



# Solution (cond.)

- The typical solutions are:
  - For event delay: Wait (Buffer).
  - For event duration: Prorate events across time slots.
- Introduce a concept of BinStream. An abstraction over the Dstream, which needs a
  - Function to extract the time fields from the records.
  - Function to prorate the records.

*Note that this can be trivially achieved by using ‘window’ functionality, by having the batch equal to the time series interval and the window size equal to maximum possible delay.*

# Problems / Solutions

- window == wait & buffer.
  - *This has two issues.*
    - A. *Need memory for buffering.*
    - B. *Downstream needs to wait for the result (or any part of it)*
- *BinStream provides two additional options*
  - *Gets rid of delay for getting partial results, by sending regular latest snapshots for the old time slots. This does not solve the memory and increases the processing load.*
  - *If the client can handle partial results, i.e. if it can aggregate partial results, it can get updates to the old bins. This reduces the memory for the spark-streaming application.*

# Limitations.

- The number of time series slots for which the updates can be generated is fixed, basically governed by the event delay characteristics.



**THANK YOU!**