DATA HANDLING PART 2

SCIANTA ANALYTICS

DEEP INSIGHT™

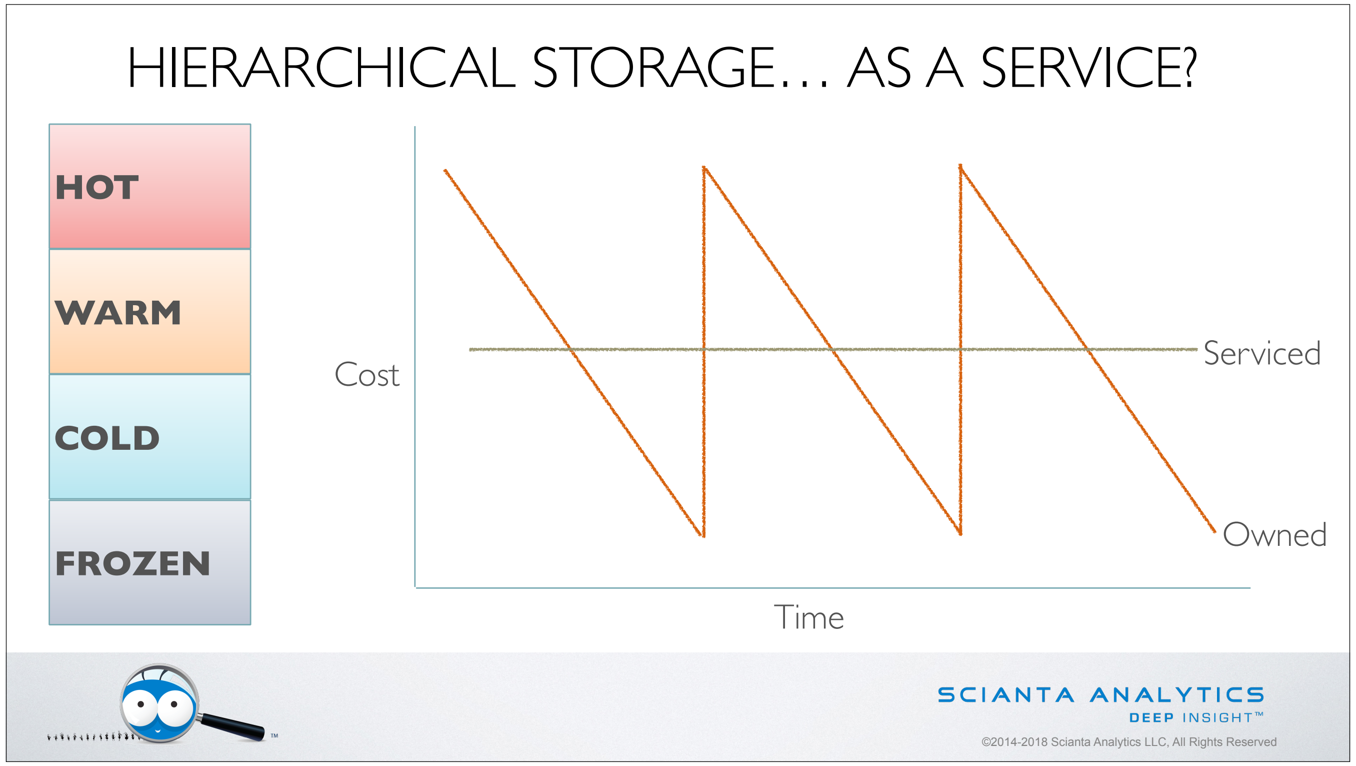
“The natural evolution of machine learning, Cognitive Computing attempts to imbue, in computer systems, the same insight and understanding we see in humans.”

Earl Cox
Chief Scientist, Scianta Analytics
Splunk .Conf 2013

AGENDA

Introduction to Machine Intelligence	Data Handling 1	Data Handling 2	Anomaly Detection	Transactional Behavior	Impact Analysis
Academic Concepts	Collection	Retention	Anomaly Definition	Defining Transactions	Organizational Visibility
Data Systems	Storage	Format	Measuring Normality	Transaction Relationships	Types of Impact
Maturity Curve	Security	Labeling		Probability Measurement	Responsiveness

SCIANTA ANALYTICS
DEEP INSIGHT™

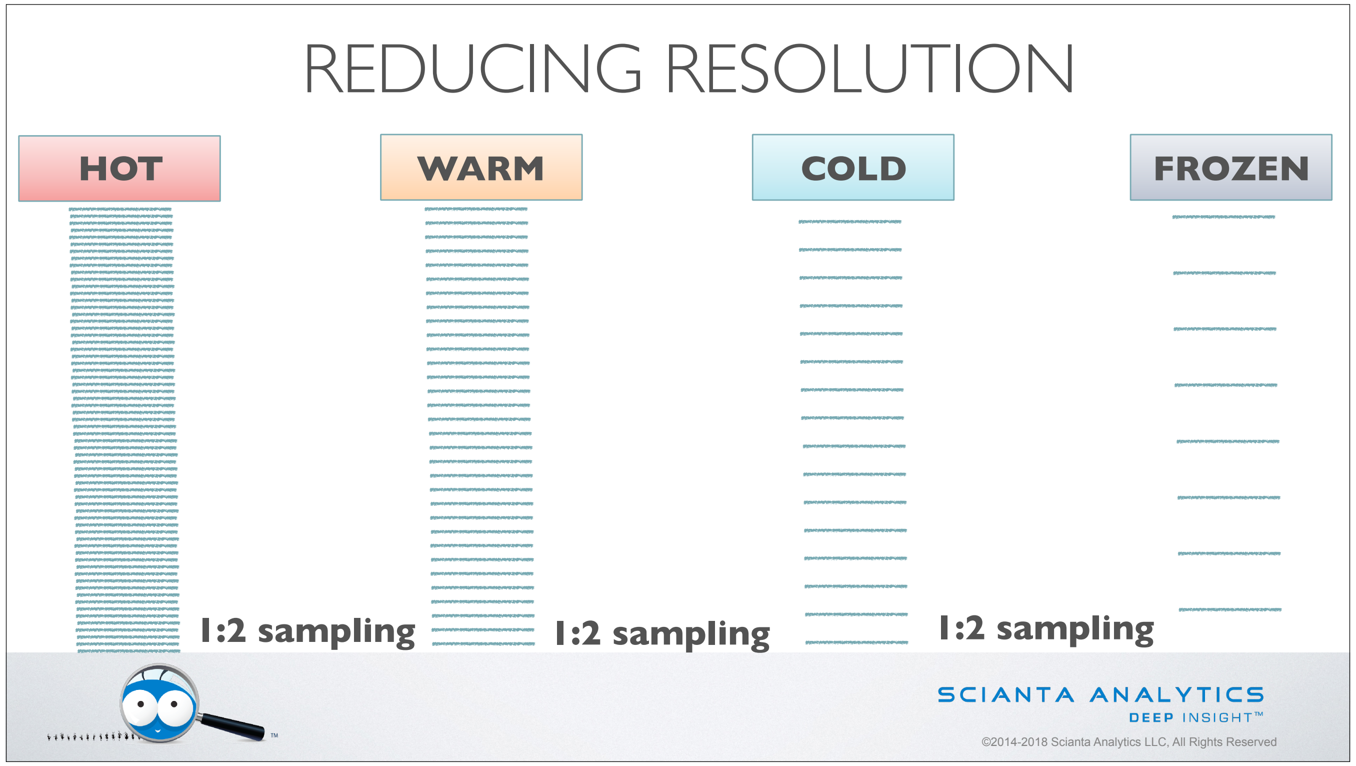


HIERARCHICAL STORAGE… AS A SERVICE?

SCIANTA ANALYTICS
DEEP INSIGHT™

In the last session we talked about the value of a hierarchical storage system; but we didn't touch on the concept of buying storage equipment versus buying storage services. Note that this doesn't only refer to providers like AWS or Microsoft. Many large organizations have internal providers with similar economics and approaches to providing basic IT services.

Theoretically, a service smooths out the spikes and troughs that can be seen in self-owned infrastructure. Where an internal system is regularly aged out and replaced with new expense, a service provider can hide these generational updates behind a single cost. However, this graph assumes that the service provider is willing to update gear at the same rate that the deployment requires, that the pricing structure doesn't change, and that the data system requirements do not change.



REDUCING RESOLUTION

HOT

WARM

COLD

FROZEN

1:2 sampling

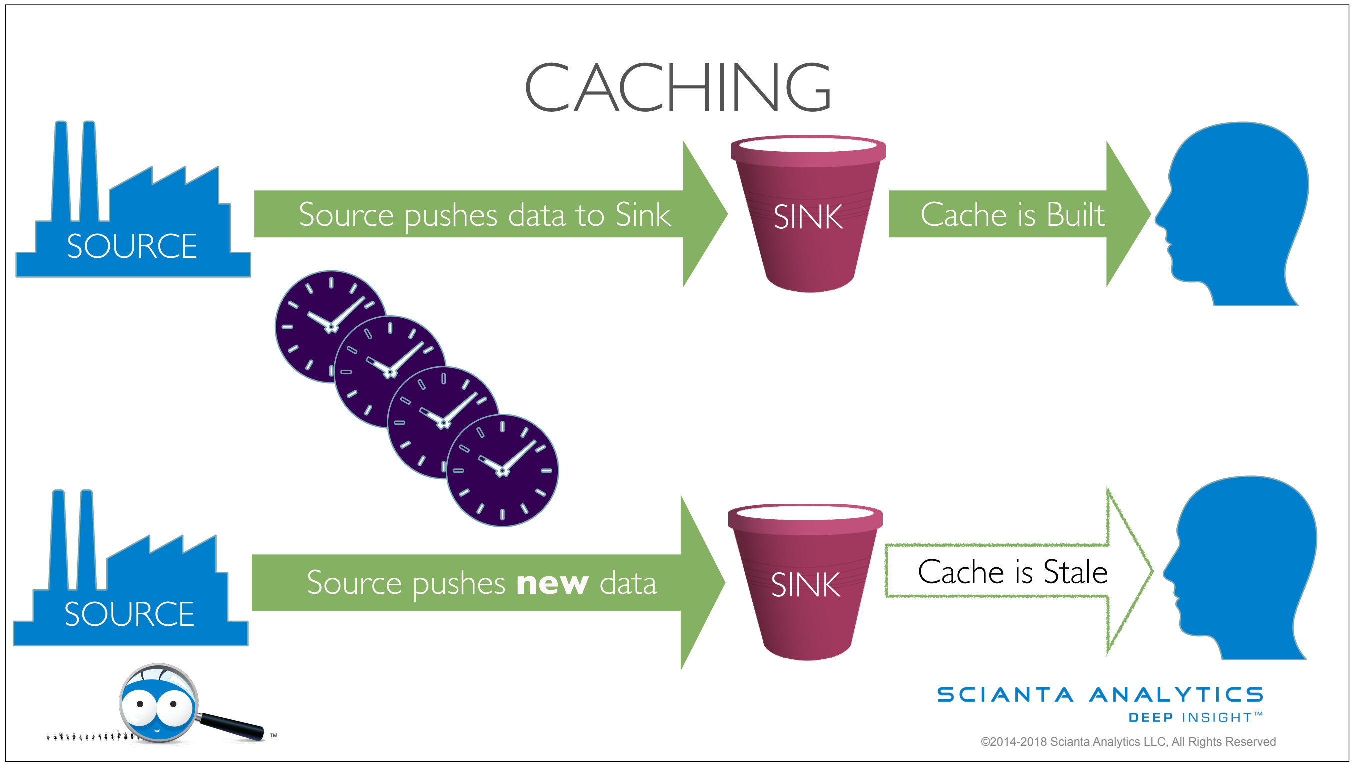
1:2 sampling

1:2 sampling

SCIANTA ANALYTICS

DEEP INSIGHT™

Some data systems encourage reducing resolution over time, particularly when designed for metrics storage. As discussed earlier this approach is not optimal for events; it also can have impact on the regulatory picture if there is an expectation of full data fidelity for the entire retention period. Finally, the data handling software being used needs to be able to perform the reduction step, which may not be easy if it wasn’t designed in by default. If all those caveats can be accepted, this approach greatly expands the timeframe of data stored in lower stages of the hierarchical storage stack.



CACHING

SOURCE

Source pushes data to Sink

SINK

Cache is Built

SOURCE

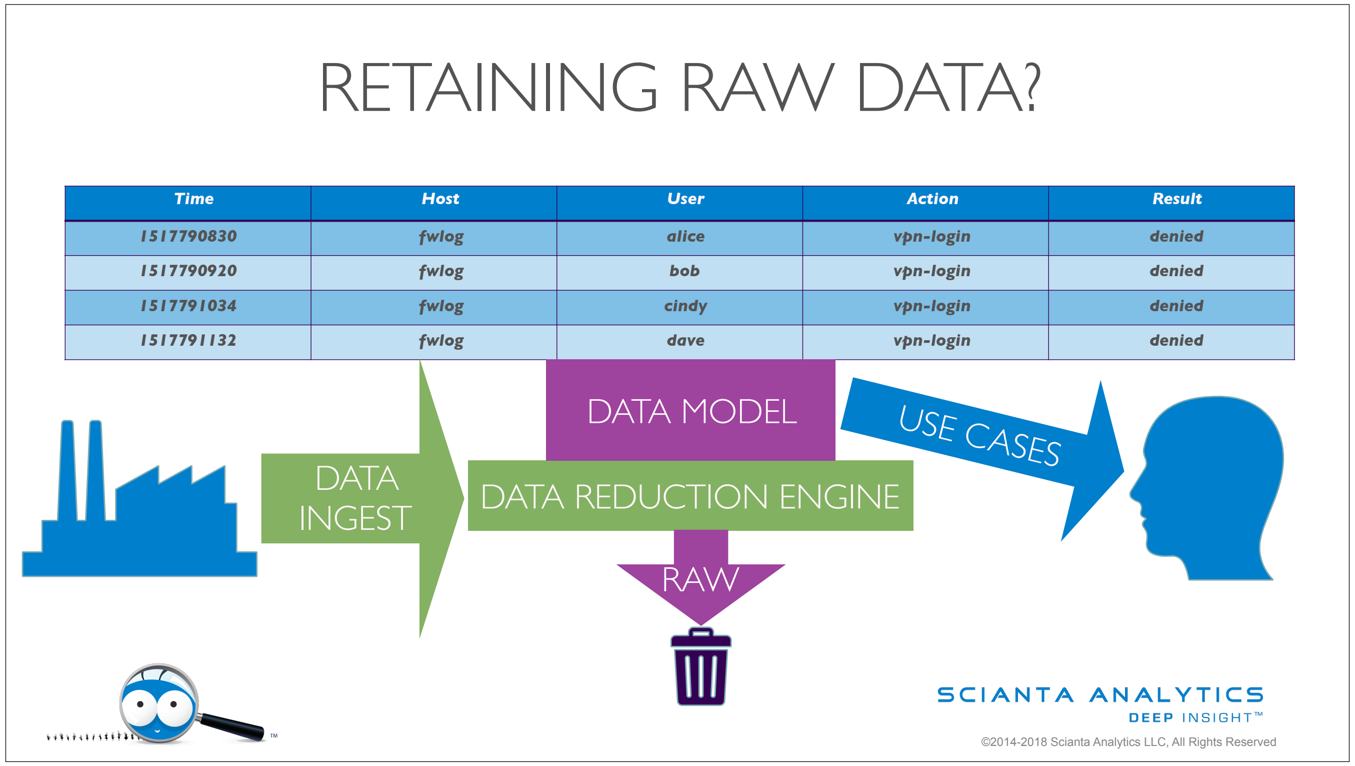
Source pushes **new** data

SINK

Cache is Stale

SCIANTA ANALYTICS
DEEP INSIGHT™

Caching data can accelerate the user's perceived performance by quite a bit. Much like the columnar data store transformation we discussed in the last session, a cache provides a faster result to the user than a direct search against raw data. However, caches can go stale and it is important to warn users that they may not be looking at the most complete data. The only way to be sure that a cache is a workable solution is to learn the user's use case. Is delay acceptable?



RETAINING RAW DATA?

Time	Host	User	Action	Result
1517790830	fwlog	alice	vpn-login	denied
1517790920	fwlog	bob	vpn-login	denied
1517791034	fwlog	cindy	vpn-login	denied
1517791132	fwlog	dave	vpn-login	denied

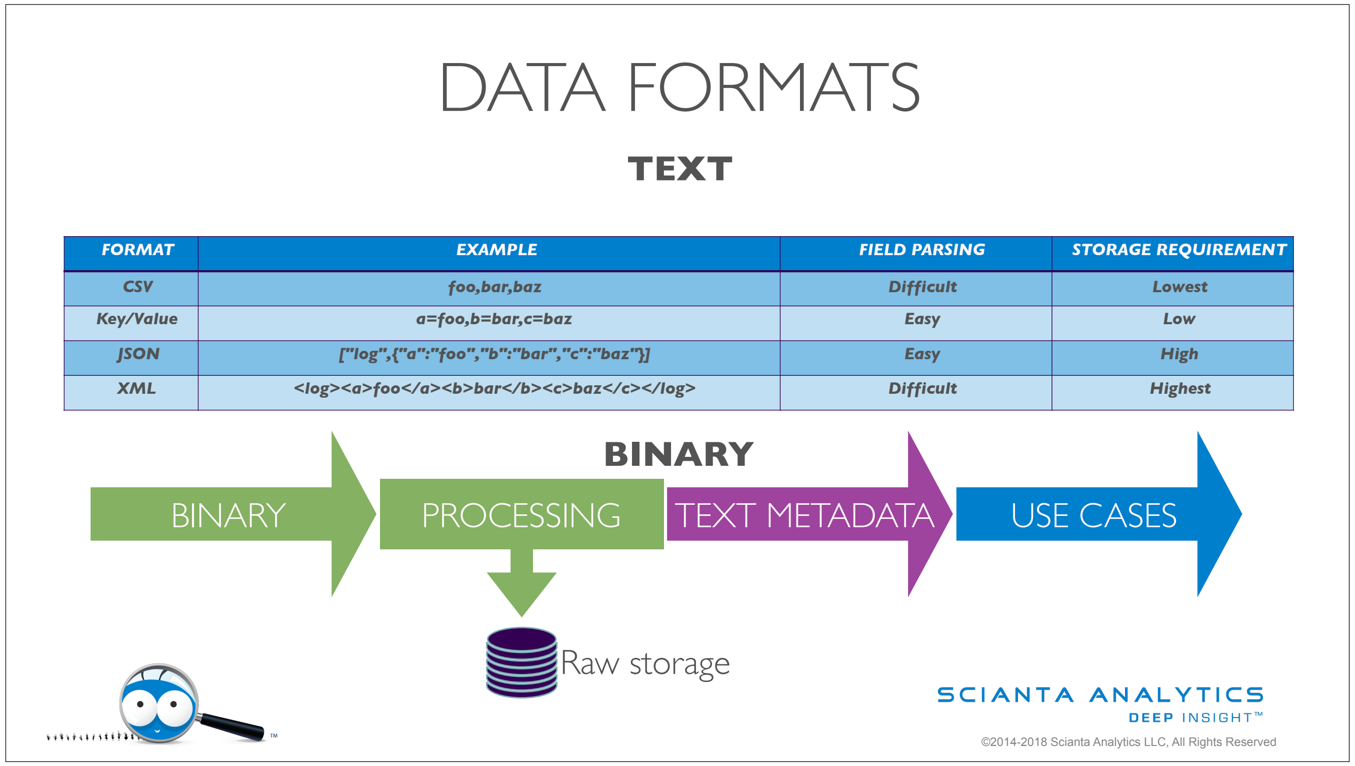
SCIANTA ANALYTICS
DEEP INSIGHT™

We've discussed sampling to reduce data and caching to increase access speed… but what if a columnar data store provides everything needed to support your users? It's a big gamble, but you can save a lot of storage space and search performance by keeping the columnar metadata and tossing the raw. A compromise approach is to direct the raw data into an inexpensive, slow storage system while keeping columnar metadata on the highest performance storage. You should certainly review these options with your data system and storage providers.

AGENDA

Introduction to Machine Intelligence	Data Handling 1	Data Handling 2	Anomaly Detection	Transactional Behavior	Impact Analysis
Academic Concepts	Collection	Retention	Anomaly Definition	Defining Transactions	Organizational Visibility
Data Systems	Storage	Format	Measuring Normality	Transaction Relationships	Types of Impact
Maturity Curve	Security	Labeling		Probability Measurement	Responsiveness

SCIANTA ANALYTICS
DEEP INSIGHT™



DATA FORMATS

TEXT

FORMAT	EXAMPLE	FIELD PARSING	STORAGE REQUIREMENT
CSV	foo,bar,baz	Difficult	Lowest
Key/Value	a=foo,b=bar,c=baz	Easy	Low
JSON	["log",{"a":"foo","b":"bar","c":"baz"}]	Easy	High
XML	<log><a>foo</a><b>bar</b><c>baz</c></log>	Difficult	Highest

BINARY

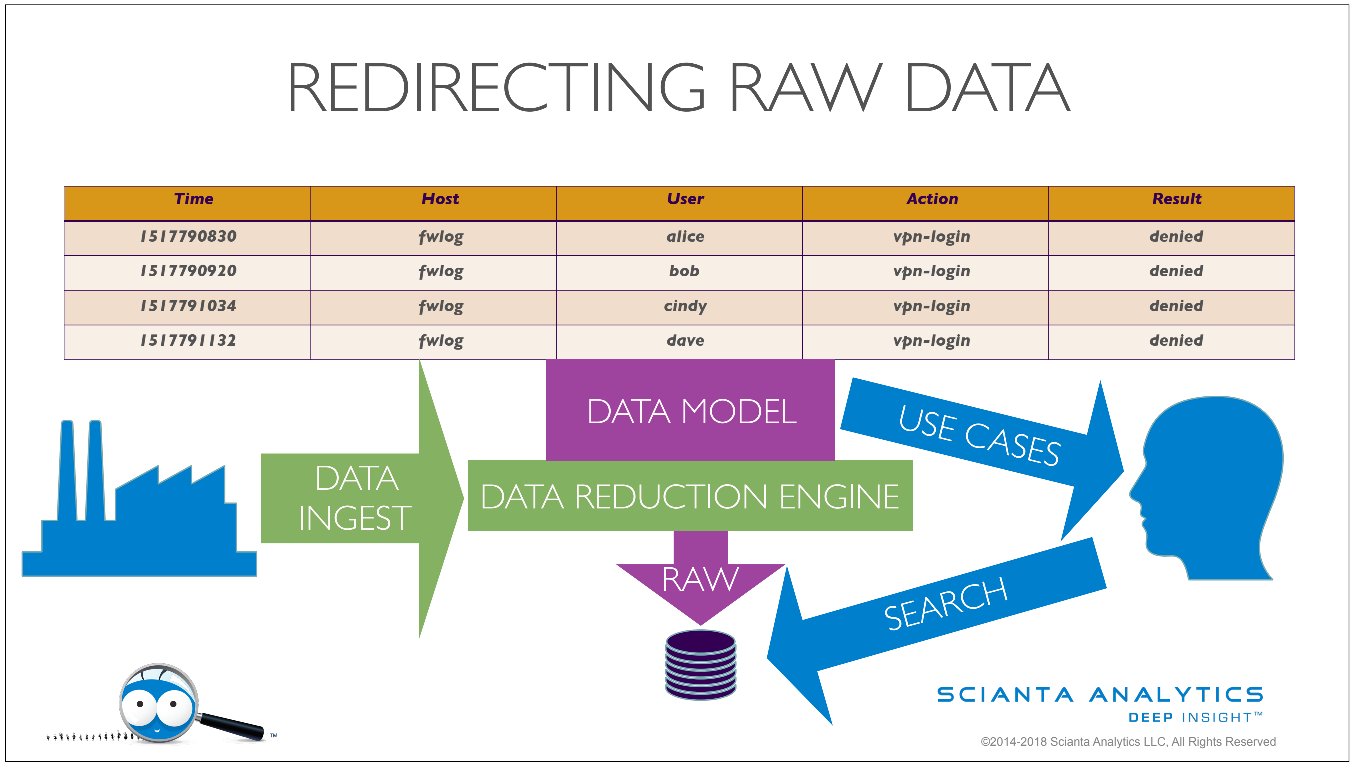
BINARY → PROCESSING → TEXT METADATA → USE CASES

PROCESSING → Raw storage

SCIANTA ANALYTICS
DEEP INSIGHT™

There's a lot of factors going into a decision about the fate of your raw data of course, and one of them is the format. Once the field values have been extracted into an columnar data acceleration or cache, this raw data format is less relevant; as long as no mistakes need to be corrected. If there's a need to return to the original analysis, demonstrate the data custody chain, or simply produce relevant raw data, that data has to be available.

Few analysis platforms are suitable for working directly with binary data; typically the data is converted to text before storage so that multiple data sources can be compared and correlated. Alternatively, a textual metadata representation can be produced from the binary, as in packet captures.

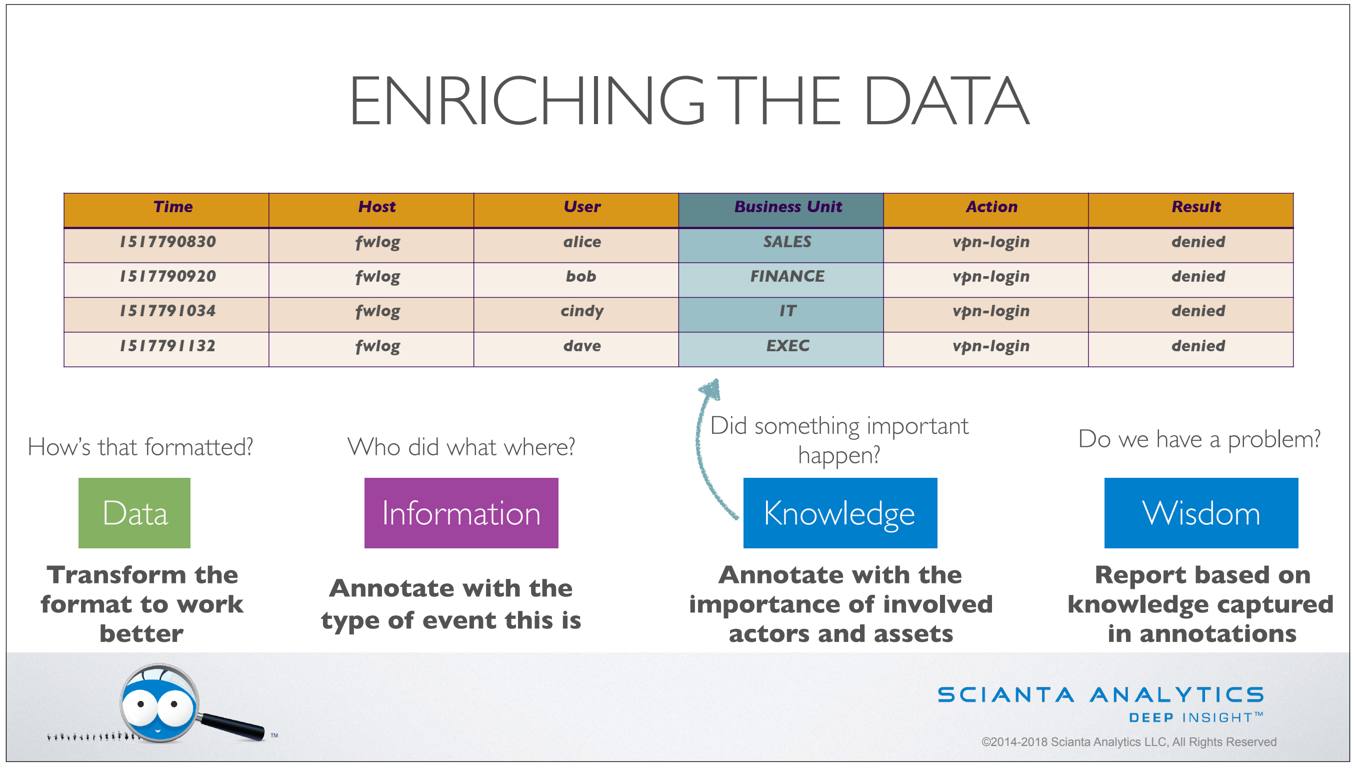


REDIRECTING RAW DATA

Time	Host	User	Action	Result
1517790830	fwlog	alice	vpn-login	denied
1517790920	fwlog	bob	vpn-login	denied
1517791034	fwlog	cindy	vpn-login	denied
1517791132	fwlog	dave	vpn-login	denied

SCIANTA ANALYTICS
DEEP INSIGHT™

So revisiting the raw data disposal idea… we've got an option here, we can redirect the raw and work with the meta. All we need to look out for is a way to find the right raw data when our meta has gotten us to something interesting.



ENRICHING THE DATA

Time	Host	User	Business Unit	Action	Result
1517790830	fwlog	alice	SALES	vpn-login	denied
1517790920	fwlog	bob	FINANCE	vpn-login	denied
1517791034	fwlog	cindy	IT	vpn-login	denied
1517791132	fwlog	dave	EXEC	vpn-login	denied

How's that formatted?

Data

Transform the format to work better

Who did what where?

Information

Annotate with the type of event this is

Did something important happen?

Knowledge

Annotate with the importance of involved actors and assets

Do we have a problem?

Wisdom

Report based on knowledge captured in annotations

SCIANTA ANALYTICS
DEEP INSIGHT™

As long as we're doing all that processing, caching, and columnar acceleration, might as well look up some values and enrich the data, right? Well, there are some catches.

First, the enriched value that you capture is what was true when you performed the enrichment. If Alice is promoted to CEO… well, your event records still says she's in Sales.

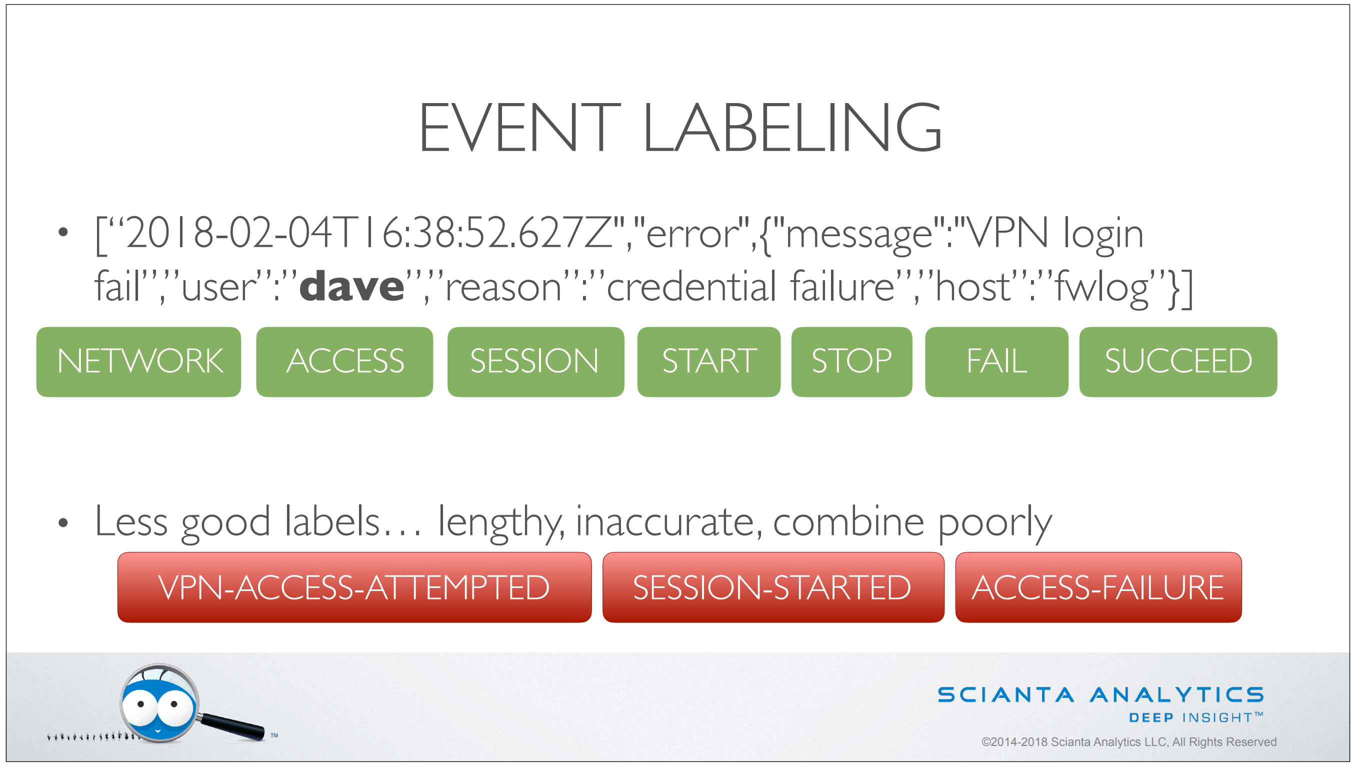
Second, that enriched value isn't in the raw data, so it's something to explain in detail to auditors and regulators. That isn't bad, it's just work to be done. So make sure you've documented the process before they come knocking.

Third, and most concerning, the enrichment process needs to be designed to go fast and fail safe, or else it can interrupt your data ingest. That's probably no big deal if your data system vendor offers the feature, but it's something to watch for when rolling your own.

AGENDA

Introduction to Machine Intelligence	Data Handling 1	Data Handling 2	Anomaly Detection	Transactional Behavior	Impact Analysis
Academic Concepts	Collection	Retention	Anomaly Definition	Defining Transactions	Organizational Visibility
Data Systems	Storage	Format	Measuring Normality	Transaction Relationships	Types of Impact
Maturity Curve	Security	Labeling		Probability Measurement	Responsiveness

SCIANTA ANALYTICS
DEEP INSIGHT™



EVENT LABELING

- [“2018-02-04T16:38:52.627Z","error",{"message":"VPN login fail”,”user”:”**dave**”,”reason”:”credential failure”,”host”:”fwlog”}]

NETWORK ACCESS SESSION START STOP FAIL SUCCEED

- Less good labels… lengthy, inaccurate, combine poorly

VPN-ACCESS-ATTEMPTED SESSION-STARTED ACCESS-FAILURE

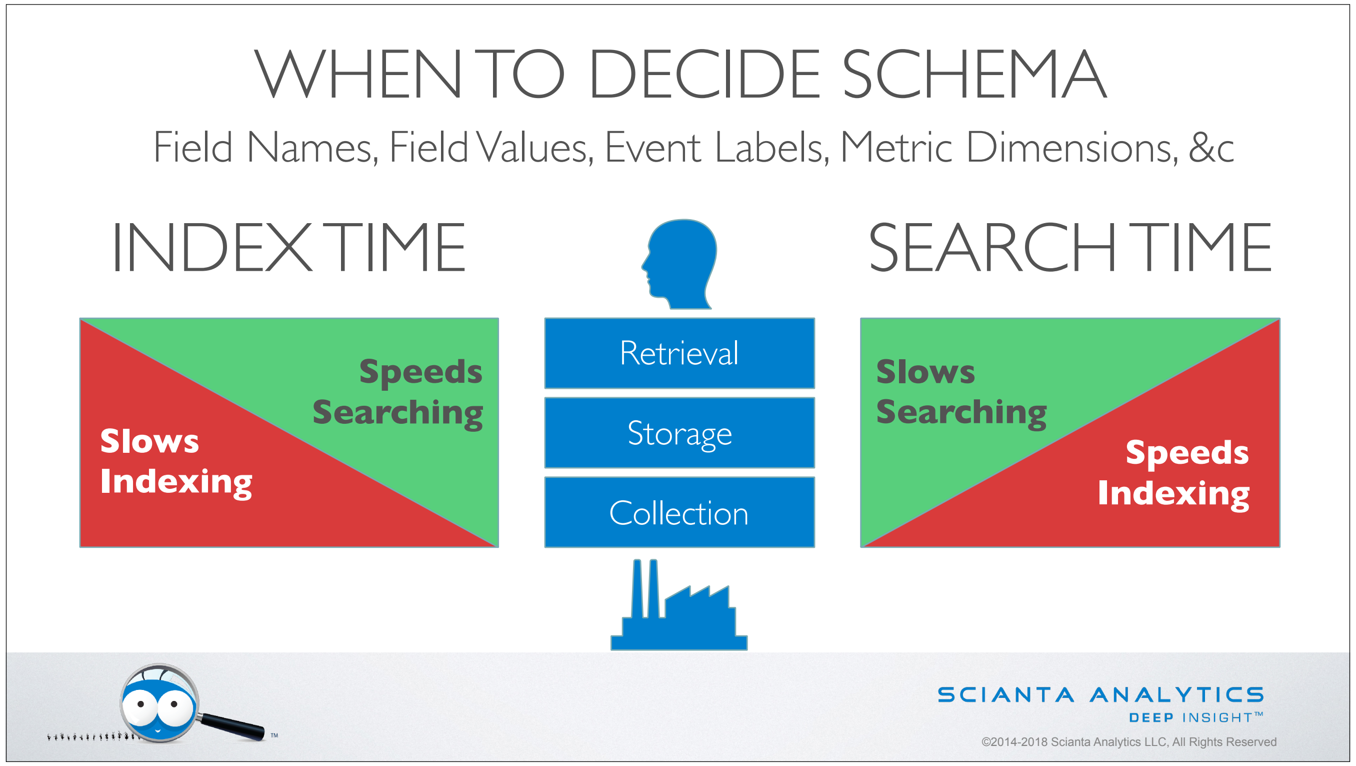
SCIANTA ANALYTICS
DEEP INSIGHT™

Depending on the system, the facility for adding metadata to raw data might be called Labels, Tags, or something else… but the concept remains the same. If you’re handed a perfect labeling schema that you just need to use, that’s great. However, you might find yourself needing to edit, extend, or create from scratch.

Use short terms that can be combined to create more complex terms.

Avoid overly specific labels that just replicate the raw event.

Encourage label-based grouping and filtering. For instance, a query for Network Session might find all VPN access attempts, while adding Start would limit the set and adding Succeed would limit it even more.

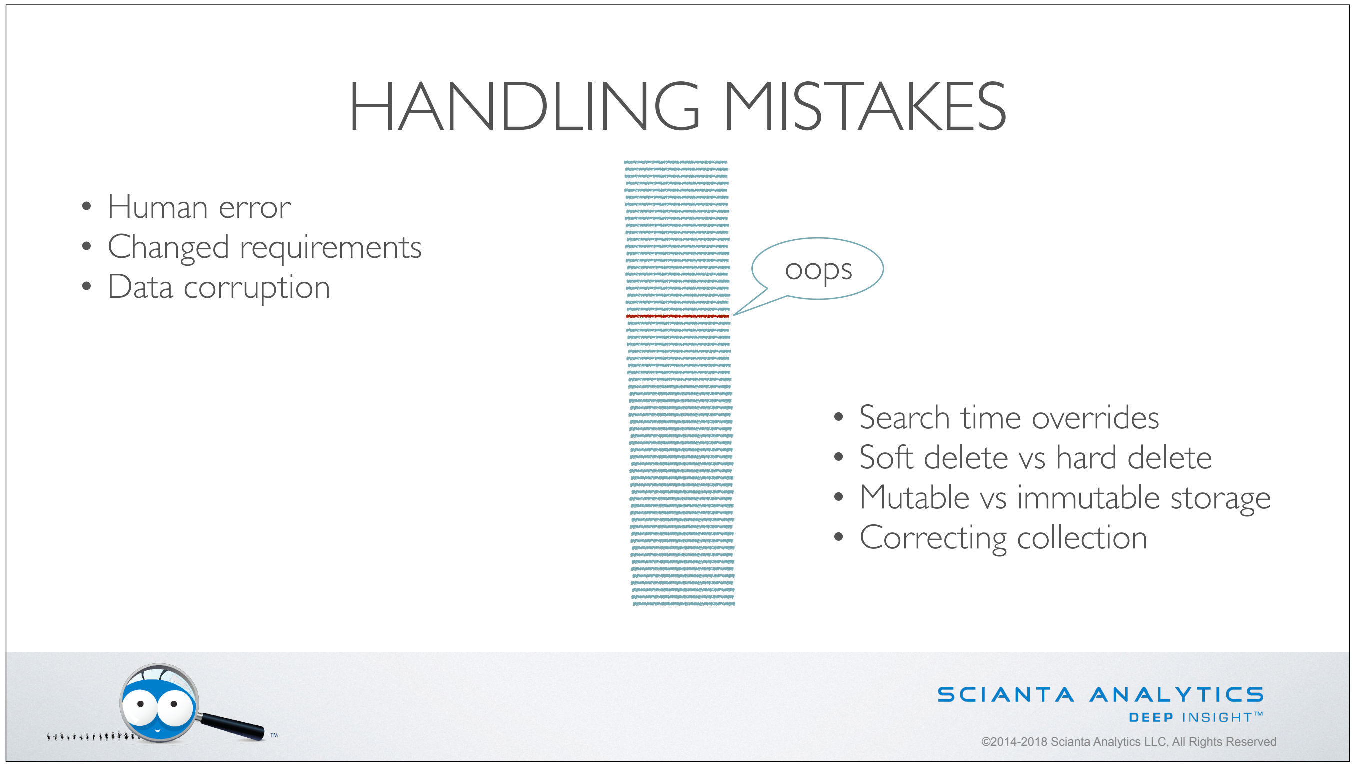


WHEN TO DECIDE SCHEMA

Field Names, Field Values, Event Labels, Metric Dimensions, &c

SCIANTA ANALYTICS
DEEP INSIGHT™

We've just spent some time looking at how to define a schema, but another factor to consider in designing a data system for cognitive computing is when the schema is applied to the data. As you might expect from any other performance-balancing endeavor, the answer depends. Which is more important, fast collection or retrieval? Does the user need to be able to modify the schema, or is this data structured properly for their use when it is created? Because those questions don't have one-size-fits-all answers, the ideal data system will allow a hybrid approach, in which the data engineer can set behavior per data type. This way the users of the system, including your cognitive modeling processes, can have their cake and eat it too.



HANDLING MISTAKES

- Human error
- Changed requirements
- Data corruption

- Search time overrides
- Soft delete vs hard delete
- Mutable vs immutable storage
- Correcting collection

SCIANTA ANALYTICS
DEEP INSIGHT™

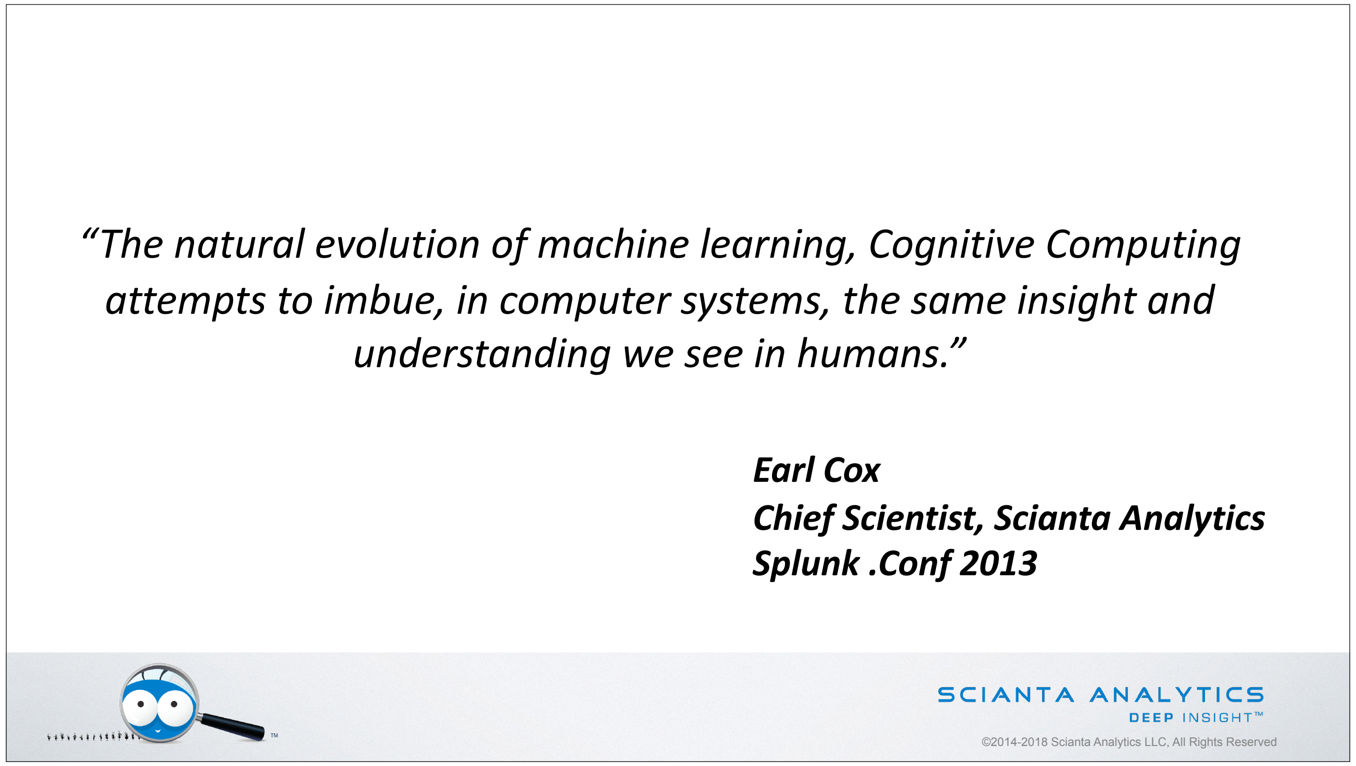
Mistakes in the data come from lots of places, and some are easier to handle than others. Ironically, uncollected data is one of the easiest to handle. Either it is possible to go back to the source and collect the data, or it isn't; either it is important enough to justify that effort, or it is not.

Far more challenging is the data that was collected, but is not collected correctly. From insidious problems like spelling errors or corrupted data to easier problems like requirements change, the data engineer is always fighting for clean data. There are four tools ready to hand in most data systems. The simplest is the search time override; regardless of whether schema was applied at index time or search time, the data engineer can apply a correction at search time and ensure that the mistake is hidden. This is suitable for things like expanded requirements — say that a numeric code is replaced with an easy to read message, for instance. However it is not suitable for collection mistakes, such as unintentional collection of data that you didn't want.

A soft delete may be temporarily acceptable in these cases; this is again a search time operation, in which the data system is instructed not to return certain records to the searching user. This is a useful temporary approach, because it hides the mistake from view until the data has aged out of retention, but bear in mind that the data is still collected and still consuming resources; it's better to make sure that you're not collecting it in the first place.

Soft deletion is not suitable for regulation failures, such as collecting personally identifiable information into a system that isn't designed to hold it. In these cases, a hard delete of the data is really the best option. You should work with your data system vendor to ensure that you understand if hard deletes are possible and how to perform a hard delete when necessary.

In other words, you're asking if the data storage is mutable, which means changeable, or immutable. There are some systems that are designed to be immutable, recording everything that they see into a permanent record. This has been variously implemented in hardware mechanisms over the



“The natural evolution of machine learning, Cognitive Computing attempts to imbue, in computer systems, the same insight and understanding we see in humans.”

Earl Cox
Chief Scientist, Scianta Analytics
Splunk .Conf 2013

SCIANTA ANALYTICS
DEEP INSIGHT™

Computers are force multipliers, not analysts; but they can help an analyst be more productive and successful. I hope this has been a useful overview of data handling; next, we will discuss using anomaly detection techniques to find interesting information in the collected data. Thank you!

Thank you!