Mapping Data Quality Concerns to Data Lake Zones

Stewart Bryson
Co-founder & CCO

twitter: @stewartbryson
medium: @stewartbryson
linkedin: stewartbryson

Stewart Bryson
Co-founder & CCO

@QualyticsData

Qualytics

The Data Quality Operating System

COMPLETENESS COVERAGE CONFORMITY CONSISTENCY PRECISION ACCURACY TIMELINESS VOLUMETRICS

qualytics.co

@QualyticsData

Qualytics is The Data Quality Operating System

Automated Profiling

Qualytics utilizes historic data to build robust metadata profiles.

This metadata is then used to infer **data quality rules**.

Anomaly Detection

Data Quality rules are notoriously hard to author and manage at scale.

Qualytics **builds** data quality rules and keeps them **up-to-date** over time.

Anomaly Remediation

The **anomaly** is the most important signal in your data pipeline.

Qualytics enables you to take **corrective actions** using your existing data tooling.

Flexible & Scalable Deployment

On-premise, single-tenant cloud or SaaS. **We meet you where your data is**.

Built on Spark and deployed via Kubernetes, **we scale with your data**.

Beyond the Basics

We can't manage what we don't **measure**.

Our **Confidence Score** exposes a qualitative measurement for every data point.

Support Modern & Legacy Data Stacks

Data Quality is important **everywhere**.

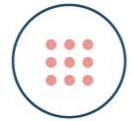
Qualytics adapts to your **data stack** and connects to anything **Spark-compatible**.

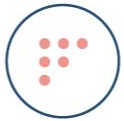
Customer Cloud or Single-tenant SaaS Deployment

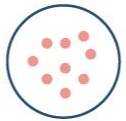
PROTECT | COMPARE | ENRICH

Supervised + Unsupervised Machine Learning

Kubernetes (Spark + Snowpark)


STRUCTURED


SEMI - STRUCTURED


UNSTRUCTURED


STREAMING

Microsoft Azure | aws | Google Cloud

@QualyticsData

There isn’t universal agreement about the *structure* and *quality* of data in the different zones.

But I will tell you what *I think.*

The difference between a Data Lake and Data Lakehouse is *not relevant* to our conversation.

The Qualytics 8

Completeness

Required fields are fully populated

Coverage

Availability and uniqueness of expected records

Conformity

Alignment of the content to the required standards, schemas, and formats

Consistency

The value is the same across all datastores within the organization

Accuracy

Your data represents the real-world values they are expected to model

Precision

Your data is the resolution that is expected

Timeliness

Data is available when expected

Volumetrics

Data has the same size and shape across similar cycles

Remediation of Data Quality issues starts with Enrichment.

Enrichment is the process of exposing anomalies and the context around them so teams can take ***corrective actions*** in data pipelines.

Pro Tip

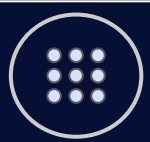
Use *inverse expectations* in Delta Live Tables to produce anomalies to enrichment tables.

Or just use Qualytics.

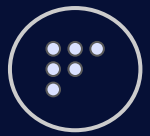
My goal in this presentation is to apply the *Qualytics 8* to the different data zones, and discuss what *enrichment* looks like in each zone.

Medallion Architecture

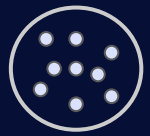
Sources



Structured



Semi-structured



Unstructured

Streaming

Zones

Bronze

Raw data, no transformations, matches source, append-only

Silver

Business entities, simplified, denormalized, standardized

Gold

Integrated, aggregated, elevated, “secret sauce”

Diamond

Published, products, applications, feature stores

Work Area

Persistent, DataOps, CI/CD, sandboxing

Raw

Transient, tables and buckets, staging and landing

Consumers

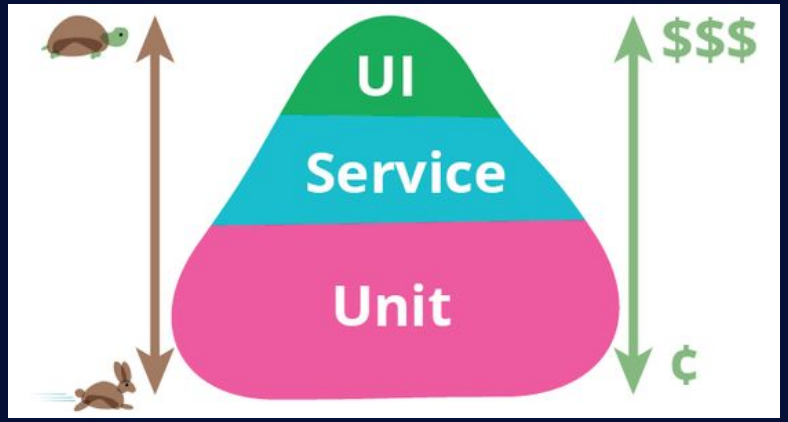
Analytics

Data Science

Ad hoc Queries

Data APIs

@QualyticsData



Fail Fast **is just as true for Data Quality checks. The cost of correcting bad data *increases* the further it moves downstream.**

Image Source: https://martinfowler.com/bliki/TestPyramid.html

Purpose

Reflect the changing state of the data in the source system.

Benefit

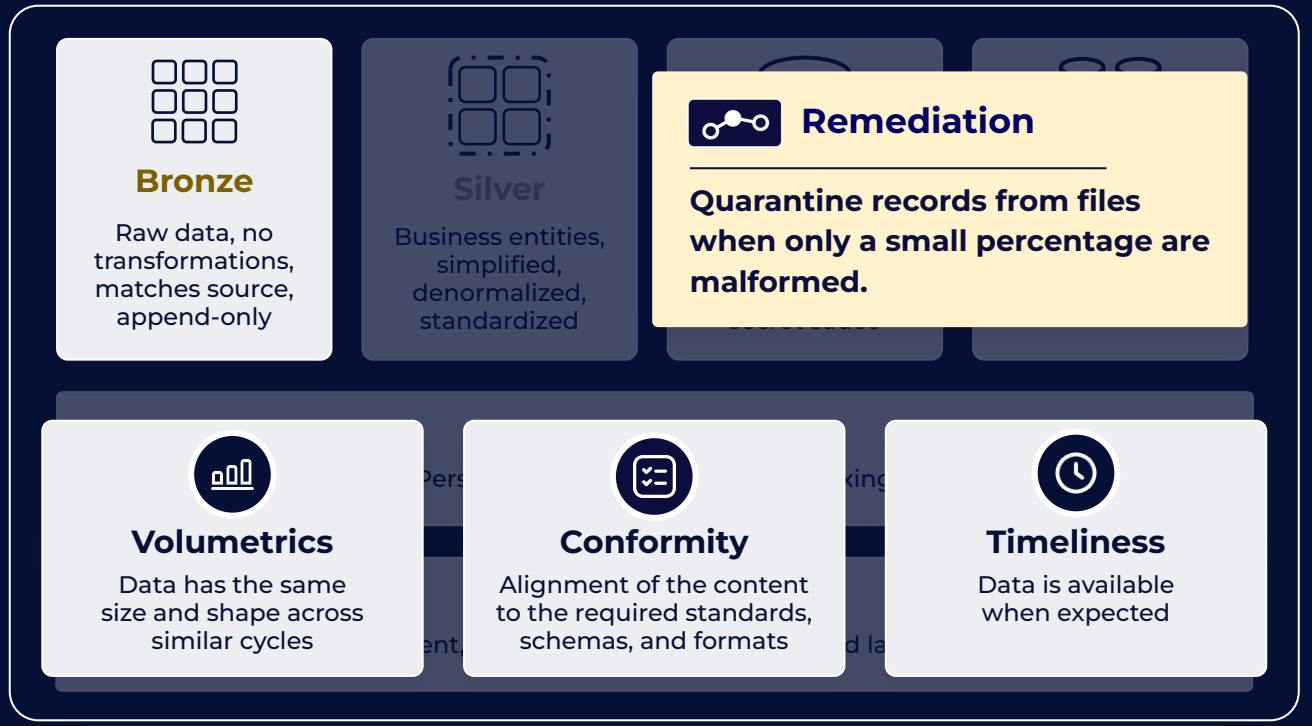
Decouple our Lakehouse from the whims of the source systems.

Purpose

Reflect the changing state of the data in the source system.

Benefit

Decouple our Lakehouse from the whims of the source systems.

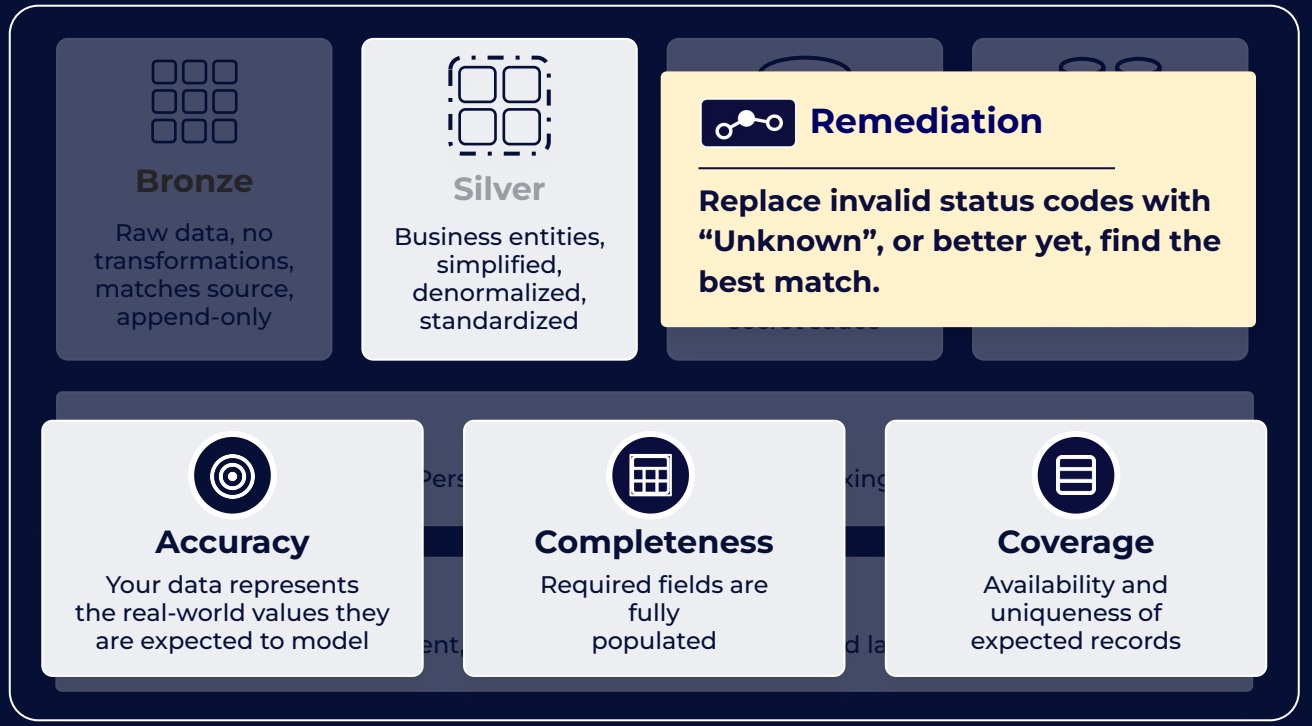


Purpose

Represent the real world events that our source systems capture.

Benefit

Reduced complexity makes this zone queryable by most analysts.



Pro Tip

The *Levenshtein Distance Formula*, available in DataFrames and SQL, is an easy way to automatically correct data in a list of values.

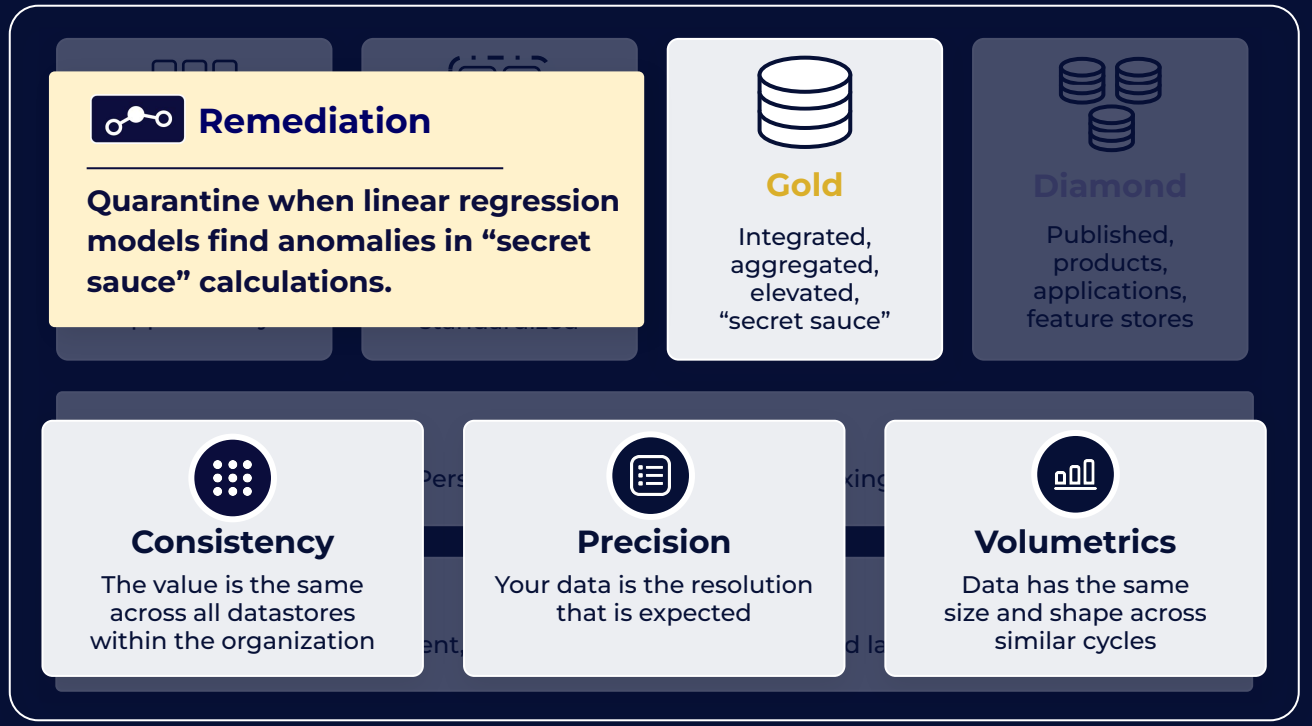
Or just use Qualytics.

Purpose

Inject opinions into models by responding to end-user requirements.

Benefit

A single version of the truth across domains, with “secret sauce”.

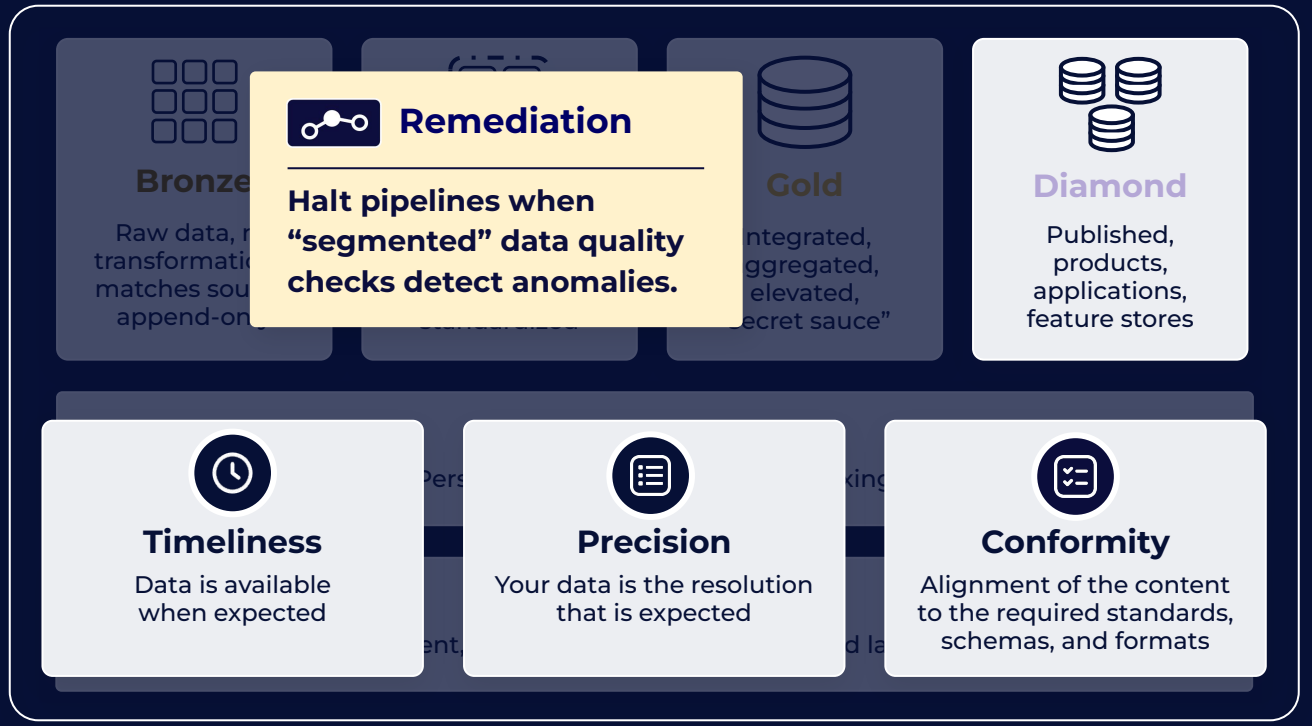


Purpose

Model the data specific to the requirements of consumers.

Benefit

Subscriptions and SLAs for data products, applications and services.



Qualytics

The Data Quality Operating System

COMPLETENESS COVERAGE CONFORMITY CONSISTENCY PRECISION ACCURACY TIMELINESS VOLUMETRICS

qualytics.co

@QualyticsData