Building an Agile Data Lake

Intelligent Data Cloud

Dale Anderson

CTO / Principal Data Solutions Architect

CDVP2

SCALEFREE

DATA
DREAMLAND
2022

AGENDA

Overview

The Agile Data Lake

ADL Implementation

Data Governance

Workflow Automation

Patterns & Practices

Metadata Driven Frameworks

Plug-n-Play Business Rules

Datapoint Repository

OVERVIEW

- *Terminology*
- *What is a Data Lake?*
- *Reference Architecture*
- *Methodology*
- *Data Modeling*

Terminology

- DaaS: *Data-as-a-Service / Data-as-a-Solution*
- Data Lake: *vs Data Warehouse vs Data-Mesh vs Data-??*
- Data Lineage: *Schematic, Semantic, Data, Process*
- Metadata: *Management & Automation*
- Business Rules: *Hard & Soft*
- Agile Data Lake: *Coupled (embedded) vs Decoupled (plug-n-play)*

What is a Data Lake?

First, let's discuss what it is NOT!

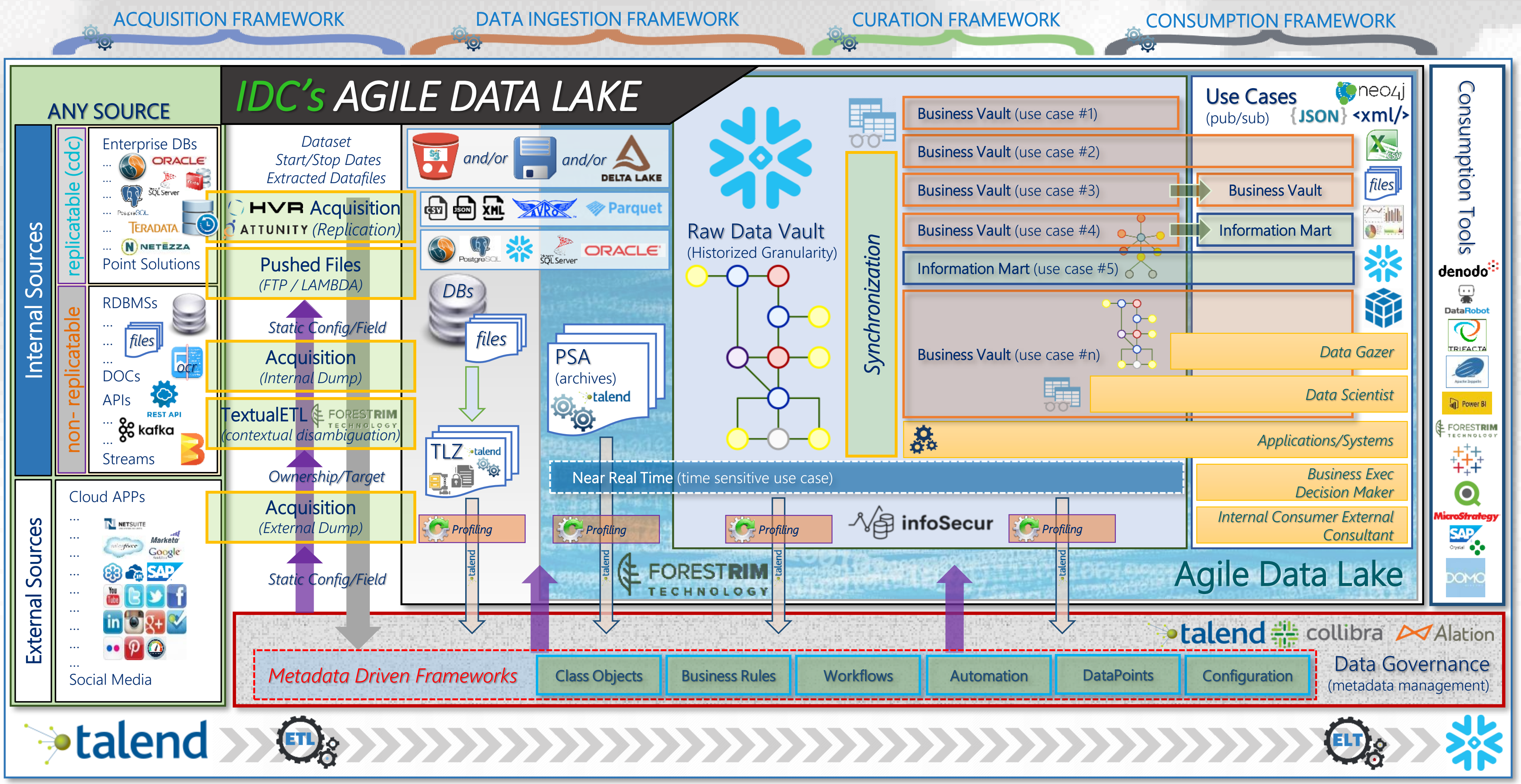
- Is NOT a new term for 'Enterprise Data Warehouse' (EDW)
- Is NOT necessarily Hadoop or NoSQL based
- Is NOT another Data Silo with fast, easy access
- Is NOT able to eliminate data integration or data processing
- Is NOT just for IoT, Analytics, & AI capabilities
- Is NOT a file store in the Cloud

What is a Data Lake?

What it IS!

- A Modernized Data Warehouse for the Enterprise
- An Architectural Strategy & Data Store for ALL business information
- A co-location of ALL Enterprise data in one virtual place
- A fusion of disparate source data with adaptable data modeling
- A well-managed, well-defined, centralized modern data warehouse
- A granular, historized, & correlated dataset for Business use
- A governed lineage data store that can be trusted

Reference Architecture



ACQUISITION FRAMEWORK

DATA INGESTION FRAMEWORK

CURATION FRAMEWORK

CONSUMPTION FRAMEWORK

ANY SOURCE

IDC's AGILE DATA LAKE

Internal Sources

replicatable (cdc)

Enterprise DBs
...
ORACLE
...
SQL Server
...
PostgreSQL
TERADATA
...
NETEZZA
Point Solutions

non- replicatable

RDBMSs
...
files
...
OCR
...
DOCs
APIs
REST API
...
kafka
...
Streams

External Sources

Cloud APPs
...
NETSUITE
...
Marketo
...
Google
...
SAP
...
...
...
...
Social Media

Dataset
Start/Stop Dates
Extracted Datafiles

HVR Acquisition
ATTUNITY (Replication)

Pushed Files
(FTP / LAMBDA)

Static Config/Field

Acquisition
(Internal Dump)

TextualETL FORESTRIM TECHNOLOGY
(contextual disambiguation)

Ownership/Target

Acquisition
(External Dump)

Static Config/Field

and/or
and/or
DELTA LAKE

CSV JSON XML AVRO Parquet

PostgreSQL SQL Server ORACLE

DBs

files

TLZ talend

PSA
(archives)
talend

Raw Data Vault
(Historized Granularity)

Synchronization

Business Vault (use case #1)

Business Vault (use case #2)

Business Vault (use case #3)

Business Vault (use case #4)

Information Mart (use case #5)

Business Vault (use case #n)

Use Cases
(pub/sub)
neo4j
{JSON} <xml/>
files

Business Vault

Information Mart

Data Gazer

Data Scientist

Applications/Systems

Business Exec
Decision Maker

Internal Consumer External
Consultant

Near Real Time (time sensitive use case)

Profiling

Profiling

Profiling

infoSecur

Profiling

talend

talend

talend

talend

FORESTRIM TECHNOLOGY

Agile Data Lake

Consumption Tools

denodo

DataRobot

TRIFACTA

Apache Zeppelin

Power BI

FORESTRIM TECHNOLOGY

MicroStrategy

SAP Crystal

DOMO

talend collibra Alation

Data Governance
(metadata management)

Metadata Driven Frameworks

Class Objects

Business Rules

Workflows

Automation

DataPoints

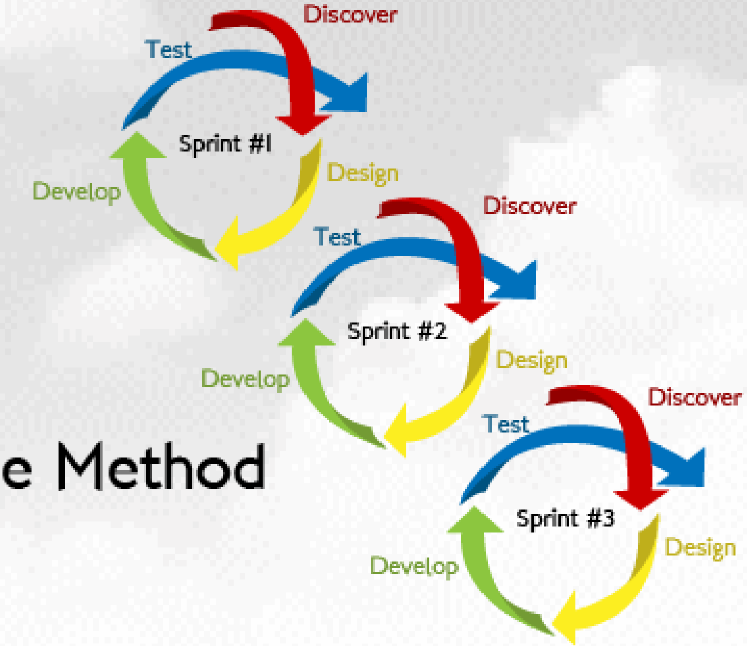
Configuration

talend

ETL

ELT

Methodology



Agile Method

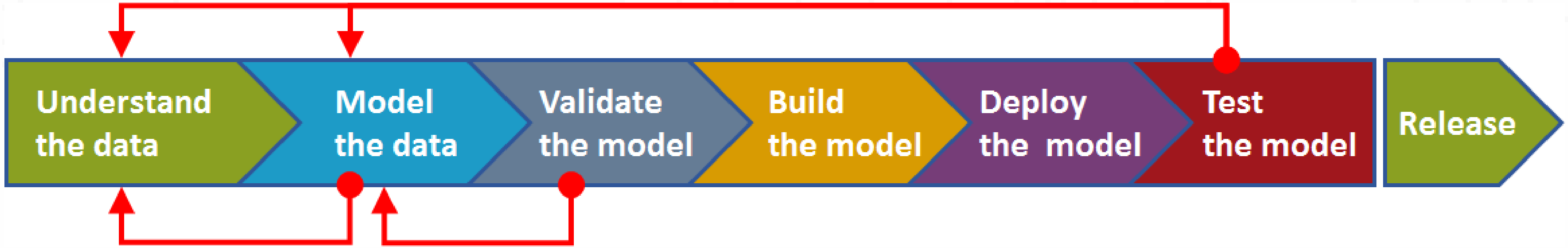
> *"Any successful methodology is one that is adopted and put into practice"*

SDLC – Software Development Life Cycle

A principled, pragmatic practice for the implementation of software applications; *establishes the foundation for software quality and incremental version release management*

DDLC – Database Development Life Cycle

A managed, prescribed practice for the design and deployment of database structures; *establishes durable, yet pliable schematic lineages for fresh installations, upgrades, and migrations*



Data Modeling

Holistic Layer

An abstract landscape of data silos across an Enterprise; *establishes the foundation for data governance*

Conceptual Layer

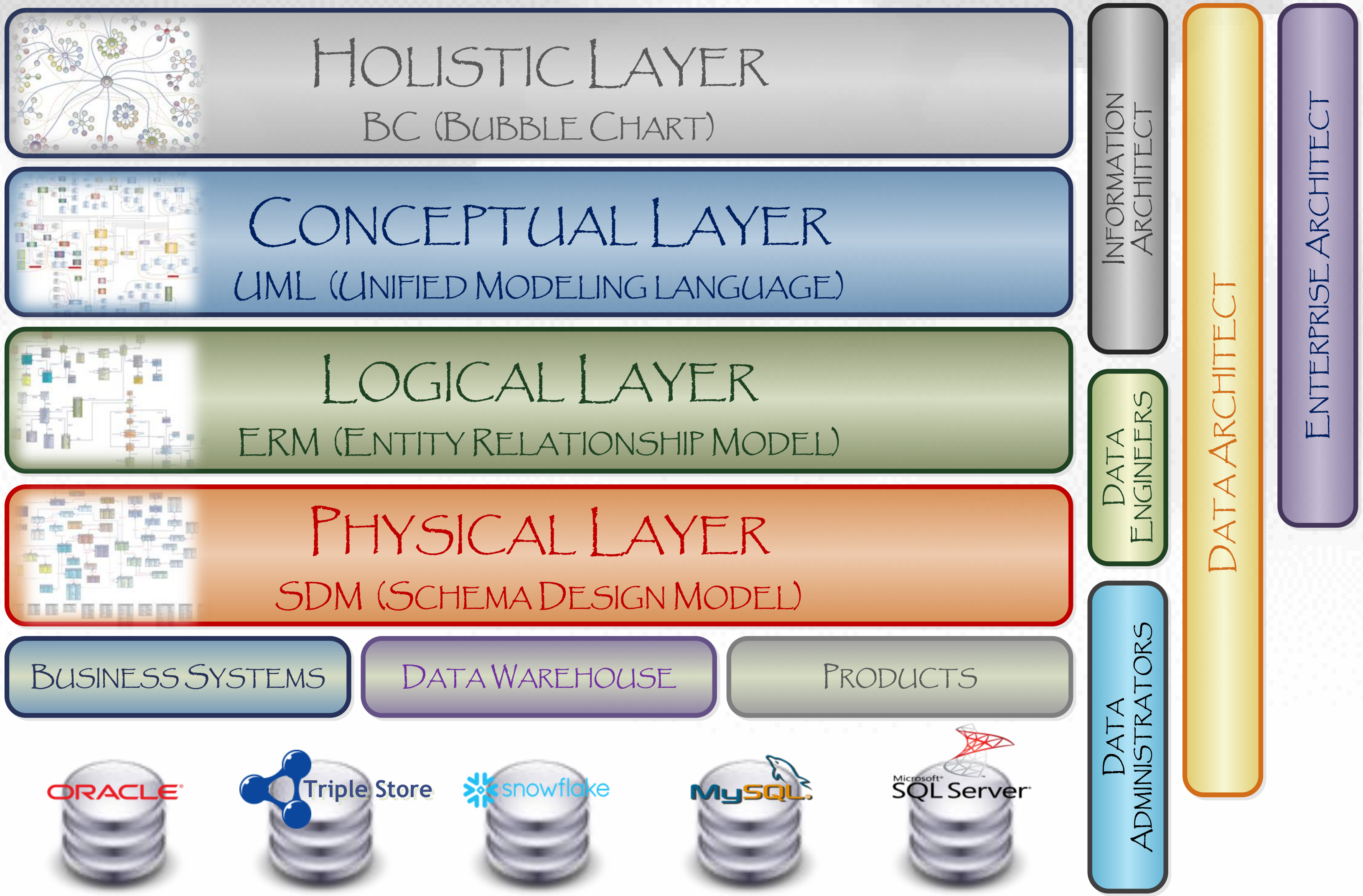
An abstract definition of business data elements and their generalized relationships; defines the semantics of the enterprise data landscape

Logical Layer

An abstract structure of semantic information organized as domain entities, attributes, and specific relationships; classifies entities, keys, attributes, and cardinality

Physical Layer

An alignment of physical artifacts with storage configurations and computing requirements; incorporates all data objects



Holistic Layer
BC (Bubble Chart)

Conceptual Layer
UML (Unified Modeling Language)

Logical Layer
ERM (Entity Relationship Model)

Physical Layer
SDM (Schema Design Model)

Business Systems

Data Warehouse

Products

Information Architect

Data Engineers

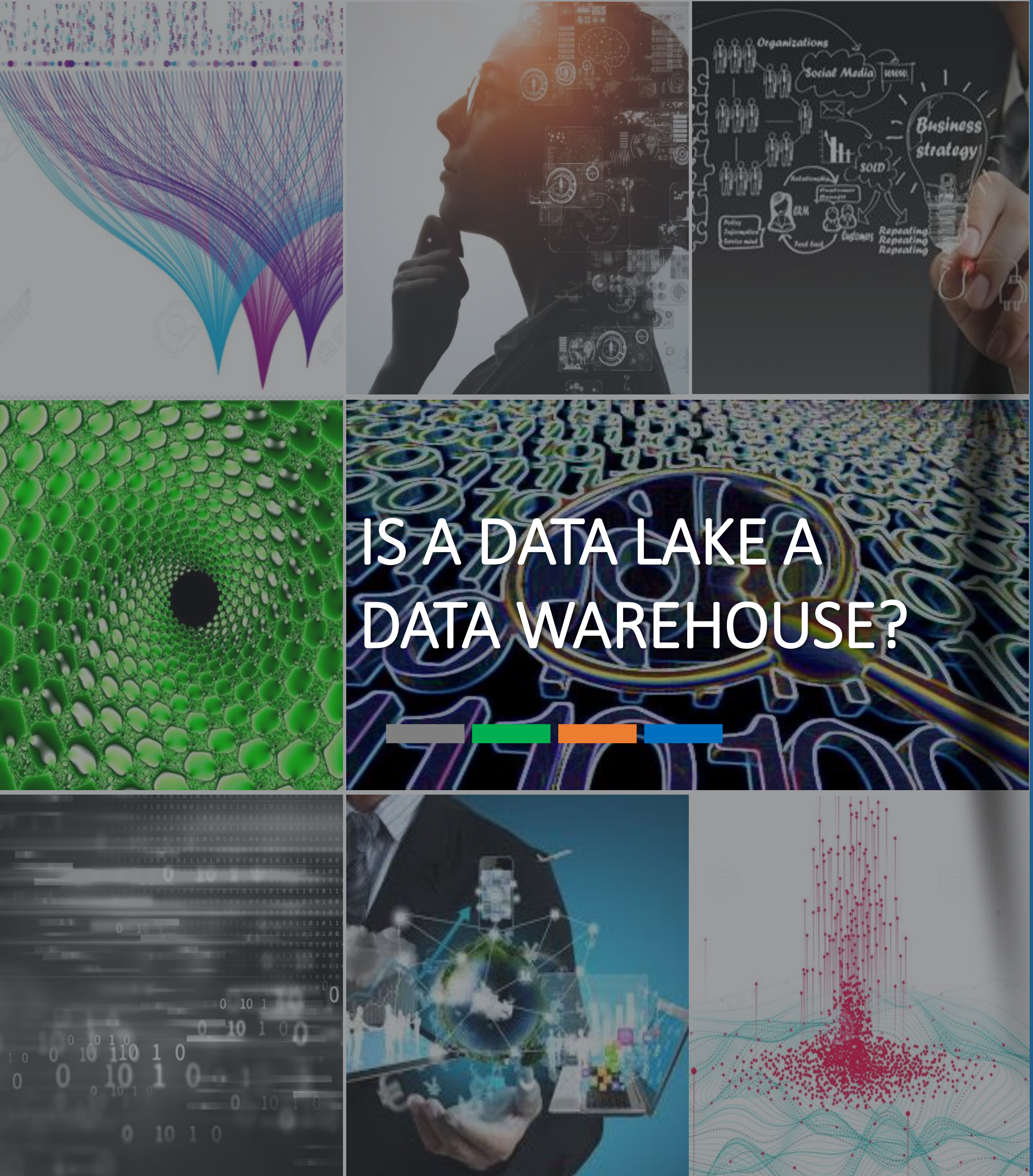
Data Administrators

Data Architect

Enterprise Architect

AGILE DATA LAKES

- *Data Lake vs Data Warehouse*
- *Why Data Vault?*
- *Agile Data Lake Process Workflows*
- *Life Cycle of an Agile Data Lake*
- *Agile Data Lake Architecture*



IS A DATA LAKE A DATA WAREHOUSE?

> "Data Lakes are a ***solution*** for Business Intelligence & Analytics – they are ***not*** a platform, they are ***not a tool,*** they are ***not a file store in the cloud!"***

Dan Linstedt

- Governed, historized, granular Data
- Business Information, Systems Integration, & Real Time data processing
- Reports, Dashboards, & Analytics
- Business Insights, Data Mining, Machine Learning, & Data Science
- 360 Visualizations
- Data Quality Monitoring/Cleansing
- Data/Systems Integration Exchange
- Customer, Vendor, Product, Service, etc.…

Why Data Vault?

”The DATA VAULT is the optimal choice for modeling the EDW in the DW 2.0 framework”

Bill Inmon

Business Alignment

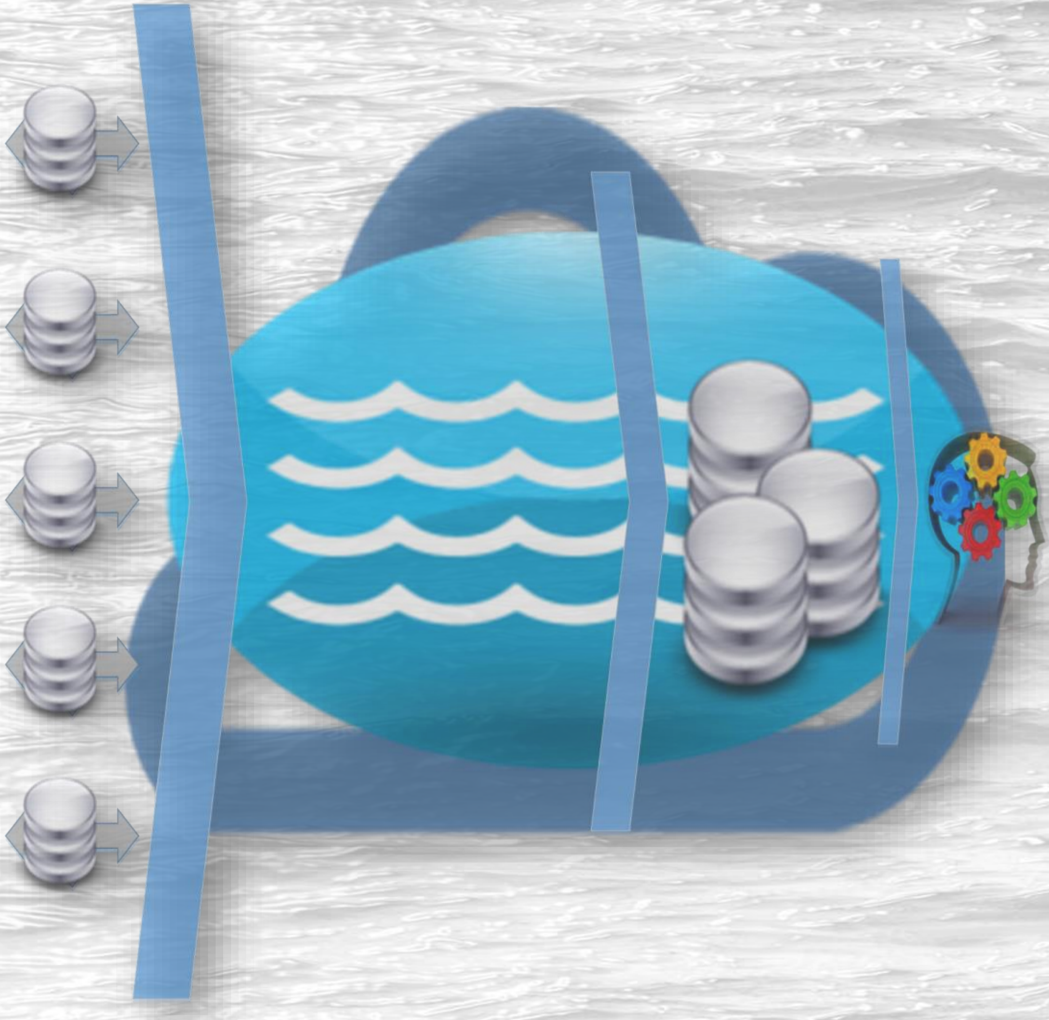
- Delivers timely, accurate business value
- Reduced design, implementation, & operational costs
- Provides auditable, historized, granular information
- Information purposed for the Business

Flexible Solution

- Well defined Architecture & Methodologies
- Incremental design, implementation, & maintenance efforts
- Adaptable Data Model
- Standards based

ADL Process Workflows

"Successful Data Lakes follow a prescribed workflow to ensure adaptability, pliability, and consistency, establishing User Trust!"



SOURCE/TARGET – Data Originations & Data Stores

Typically, Database Systems, Data Files, API Result Sets, and/or Data Streams; *essentially to understand the size and shape of SOURCE data originations is to know what and how data must be processed into TARGET data stores*

VOLUME, VARIETY, VELOCITY!

Life Cycle of an Agile Data Lake

SCALEFREE
DATA DREAMLAND 2022

Data Acquisition Workflow

Extracting raw source data, accumulating in a transient landing zone for downstream processing

Data Ingestion Workflow

Directing landing zone data objects to a persistent staging area & processing further into a raw data store (data vault)

Data Curation Workflow

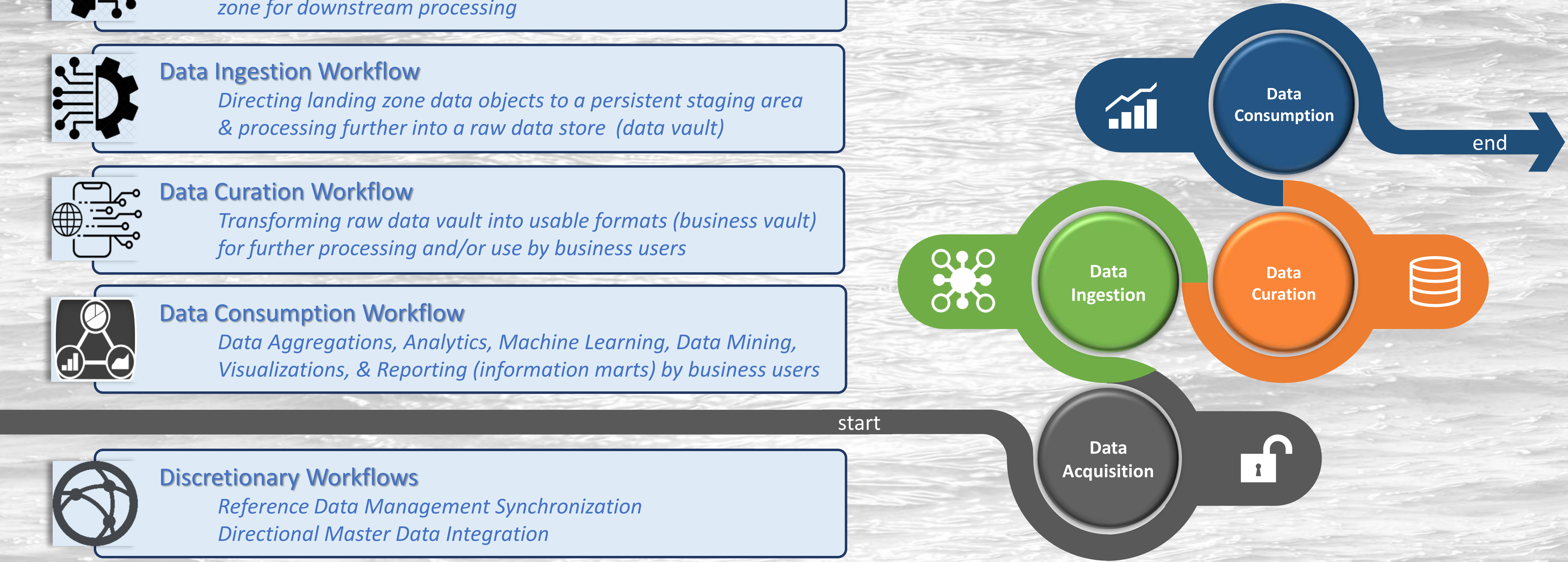
Transforming raw data vault into usable formats (business vault) for further processing and/or use by business users

Data Consumption Workflow

Data Aggregations, Analytics, Machine Learning, Data Mining, Visualizations, & Reporting (information marts) by business users

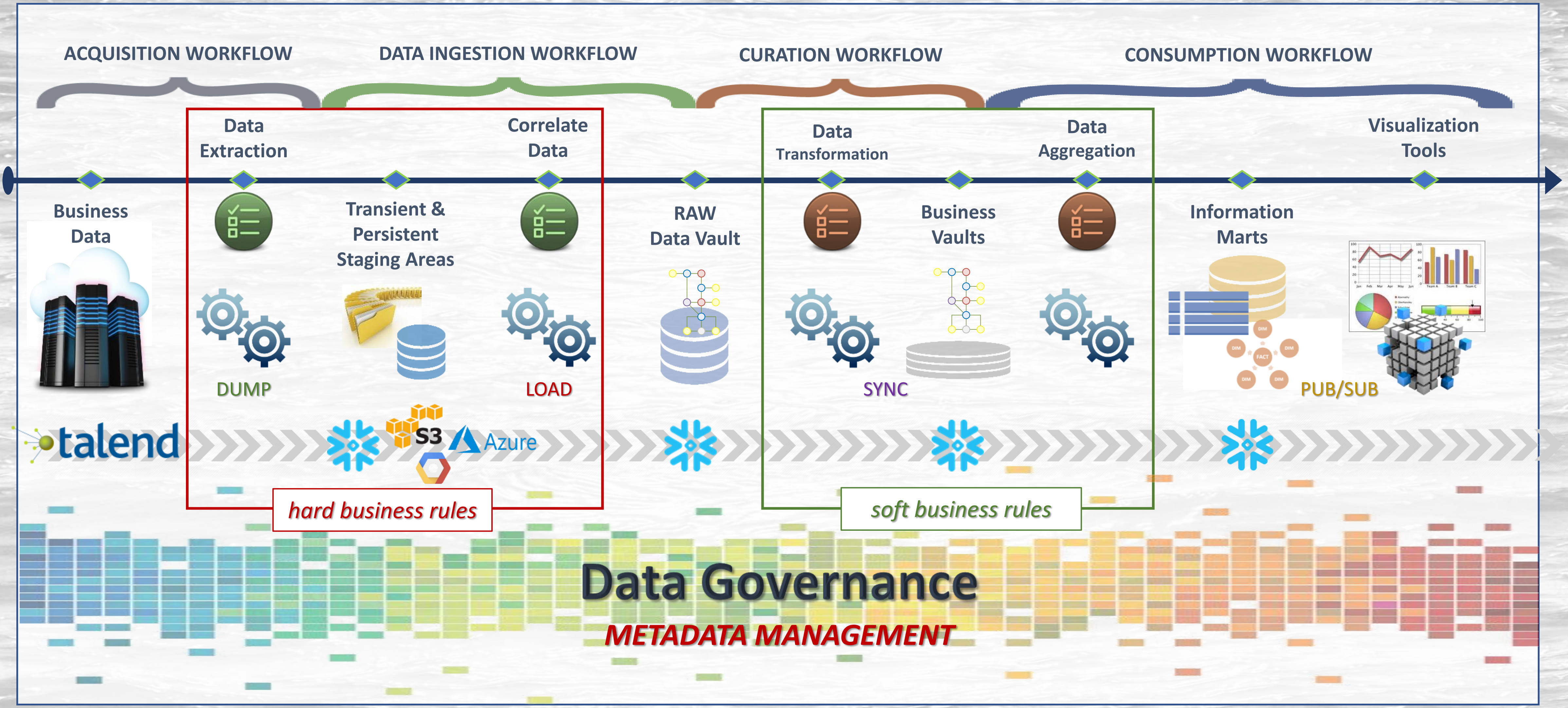
Discretionary Workflows

Reference Data Management Synchronization
Directional Master Data Integration



ADL Architecture

SCALEFREE
DATA DREAMLAND 2022



ADL IMPLEMENTATION

- *Delivering Intelligent Data*
- *Technology Integration*
- *Project Management*

Delivering Intelligent Data

SOFTWARE

- Licensing
- Design
- Development
- Automation
- Support
- Tools

HOSTING

- Cloud/On-Premises
- Storage
- Compute
- Networking

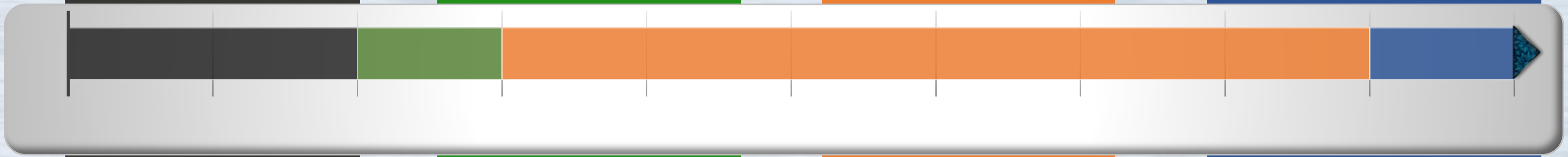
RESOURCES

- Stakeholder
- Implementation
- Governance
- Project Management
- Training

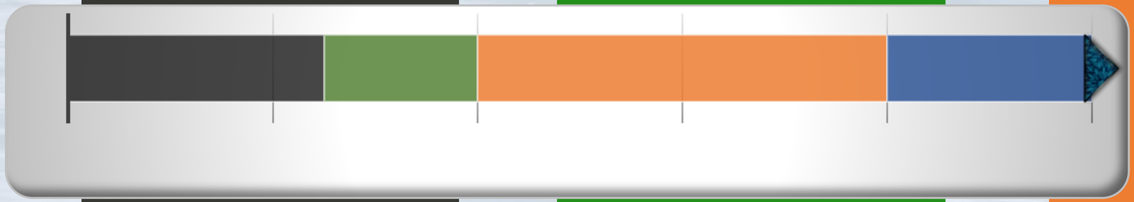
MAINTENANCE

- DevOps
- Performance Tuning
- Enhancements
- Defect Corrections

YOU COULD *BUILD* it from scratch



OR *BUY* it then customize



Intelligent Data Cloud

Technology Integration

SCALEFREE
DATA DREAMLAND 2022

Ecosystem

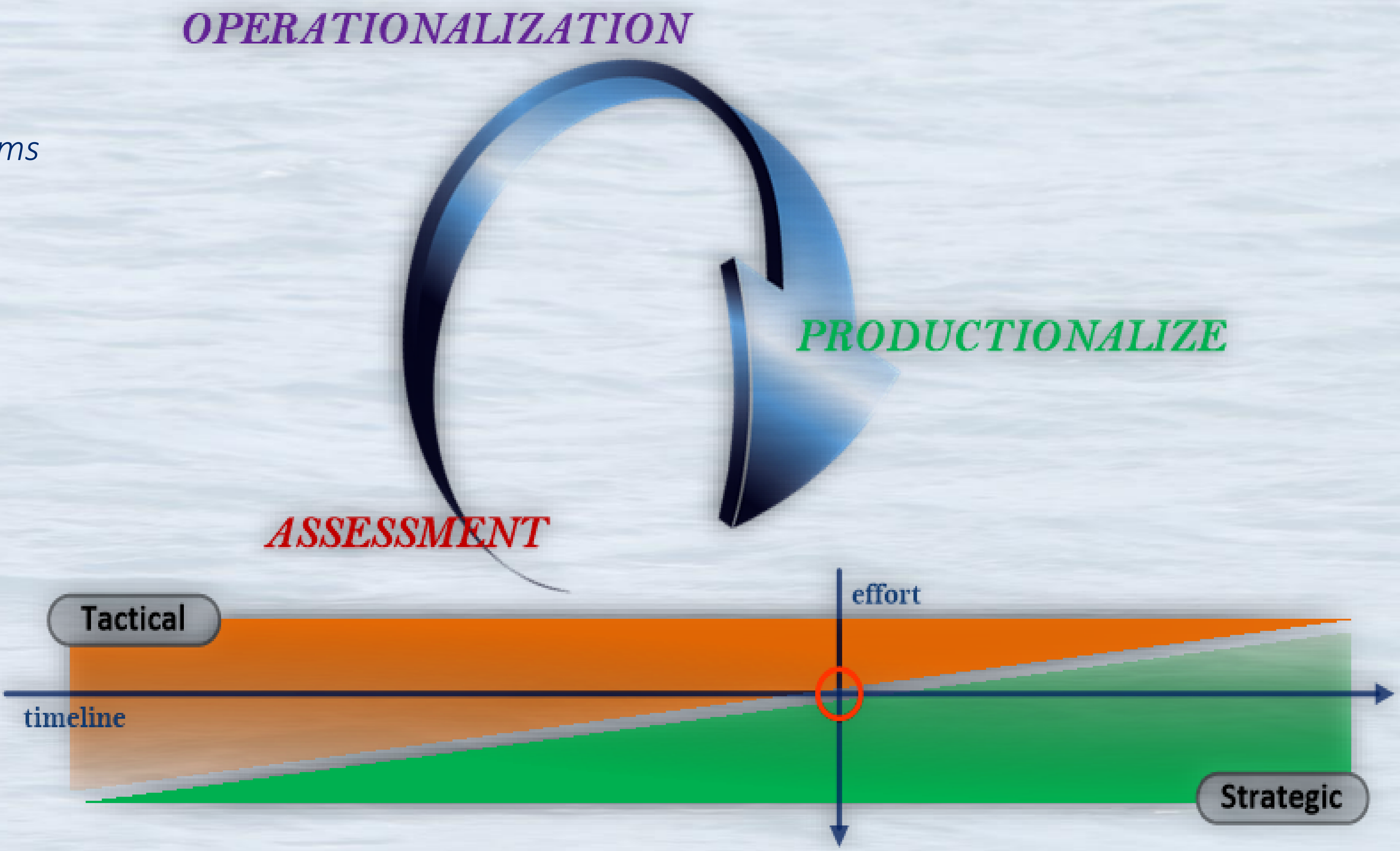
Environment:	*DEV/TEST/PROD*
Platform:	*On-Premises / Cloud*
Software:	*Tools / Engines / Systems*
Data Stores:	*Database / File Systems / APIs / Streams*
Network:	*Virtual Machine / RDP / Host Systems*
Security:	*Role Based Access Control / MFA*

Design & Implementation

Modeling:	*Process & Data Workflows*
Guidelines:	*Development Best Practices*
SDLC / DDLC:	*Versioning / Branch & Tags*
ETL / ELT:	*Hand Code / Code Generators*
Context:	*Configuration Variables*
Exceptions:	*INFO / WARN / ERROR / FATAL*
Logging:	*Return Codes / Alerts / Notifications*

DevOps

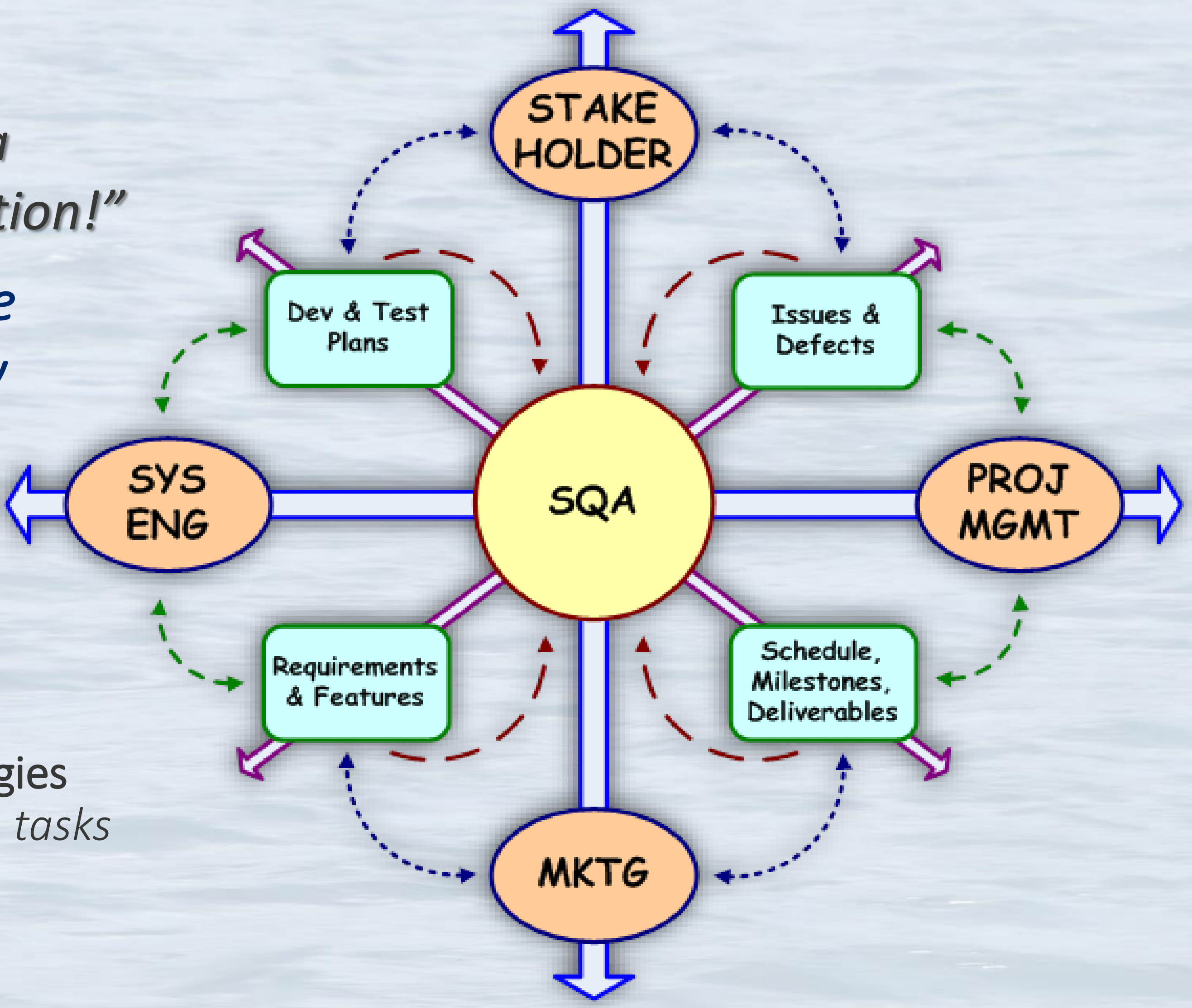
Artifacts:	*Builds / Release Management*
Engines:	*Runtime Execution Servers*
Scheduler:	*Task Operations & Monitoring*



Project Management

"Successful Data Engineering Projects need a well-balanced process and clear communication!"

- *Insufficient communication leads to failure*
- *too much process leads to perpetual delay*



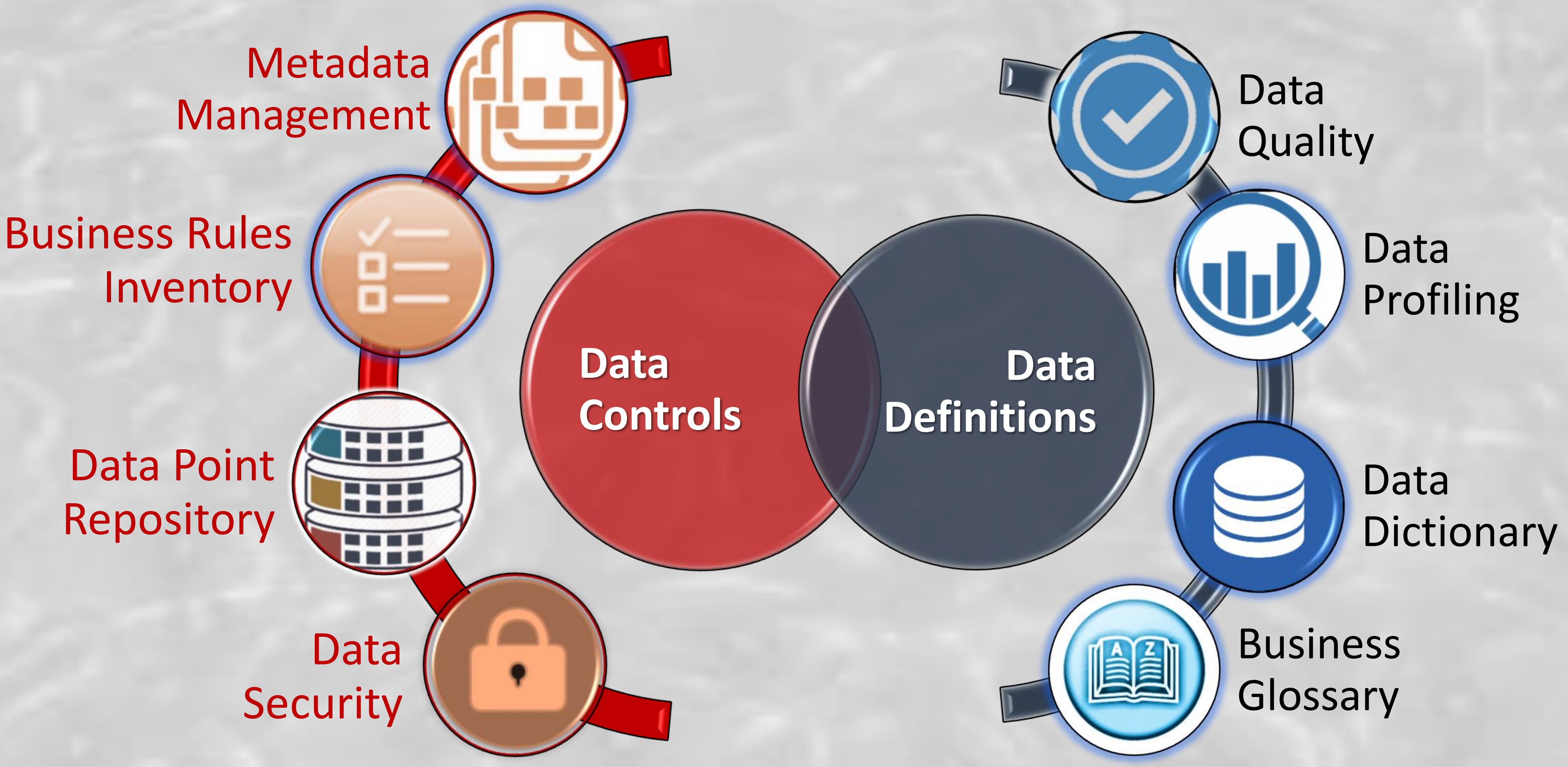
JEP – Just Enough Process

Centered around Quality, JEP codifies Agile methodologies with project disciplines; *focus is on what, how, & when tasks or stories should be performed across sprints*

DATA GOVERNANCE

- *What is Data Governance?*
- *Business Glossary & Data Dictionary*
- *Taxonomy of a Business*
- *Data Quality & Stewardship*
- *Metadata Management*

What is DATA GOVERNANCE?



Metadata Management

Business Rules Inventory

Data Point Repository

Data Security

Data Controls

Data Definitions

Data Quality

Data Profiling

Data Dictionary

Business Glossary

Business Glossary & Data Dictionary

"Defining a business is like trying to explain how an airplane flies; each element of flight can be described in many ways, and often is…"

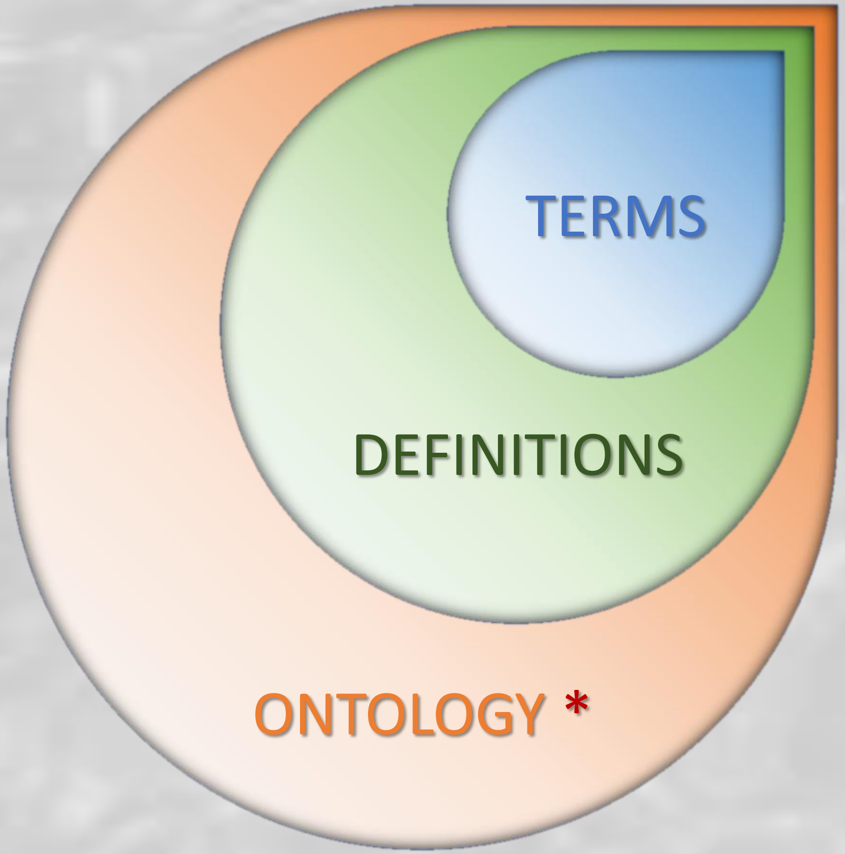
Business GLOSSARY – The Taxonomy of a Business

Clear, concise, accepted definitions of business terminology; *essentially to understand the business at its core enables all who participate in its success – semantic lineage*

Data DICTIONARY – The Value of Data Structures

Factual, historical, data structures from business data systems; *essentially to understand the structure of SOURCE & TARGET data and their mappings enables trusted data processing – schematic lineage*

Taxonomy of a Business



Ontology – The Information Model of a Business

Conceptual organization of the Business by Domain; *essentially a hierarchal and/or composite abstraction, grouping concepts with relationships and/or associations*

Terms – The Language of the Business

Terminology used by the Business; *essentially a collection of words, compound words, and/or acronyms that are essential to the corporate vernacular*

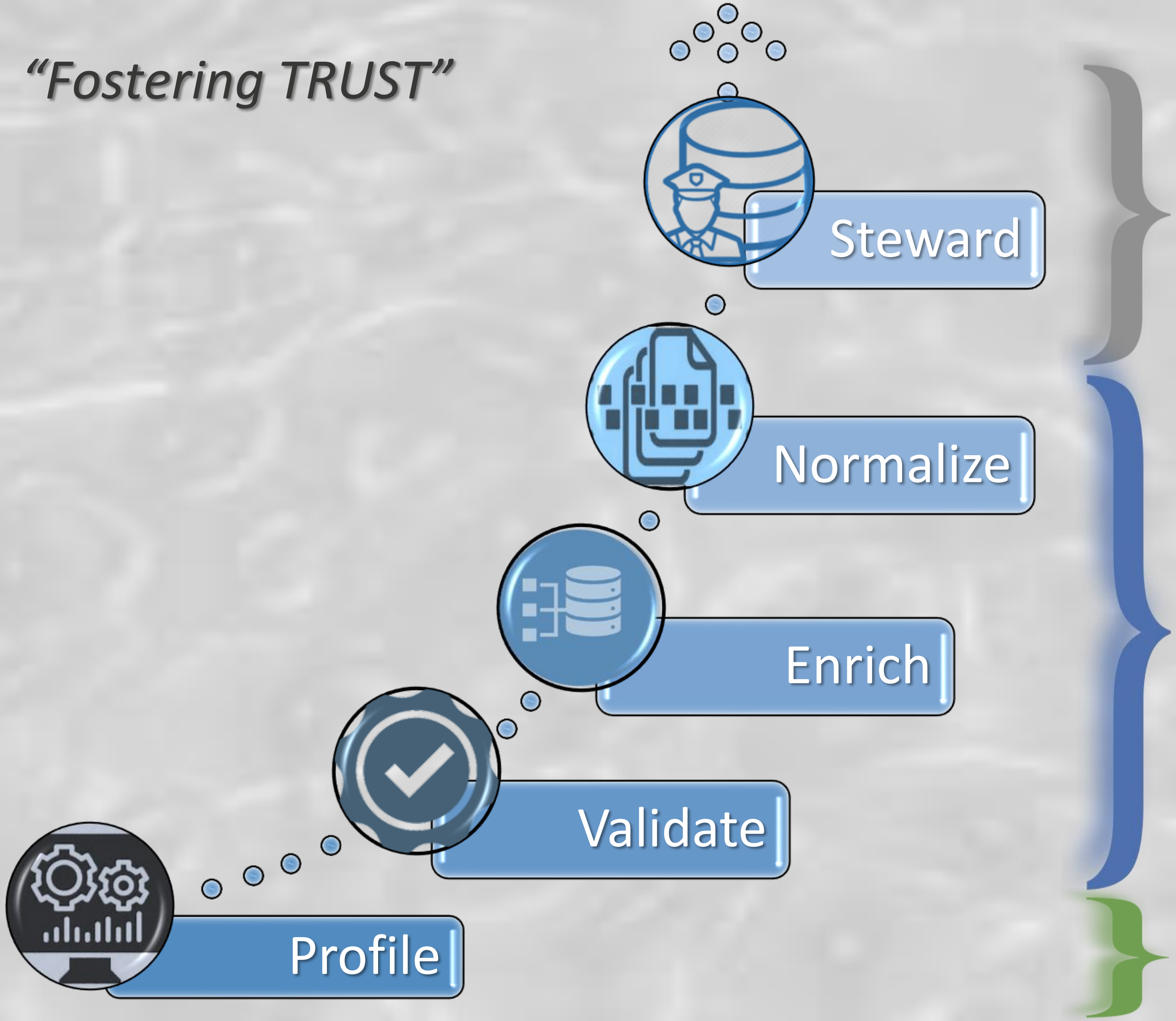
Definitions – The Meaning of Business Terminology

Widely accepted explanation of each Business Term; *essentially each identified term needs a clear description of its meaning that business stakeholders can agree on*

**** a Term may have multiple definitions depending upon the business context***

Data Quality & Stewardship

"Fostering TRUST"



Data STEWARDSHIP – Manual Process

Typically, any rejected data processed by Business Rules; *when automation is not enough manual intervention maybe required*

TRANSFORMATIONS – Automated Process

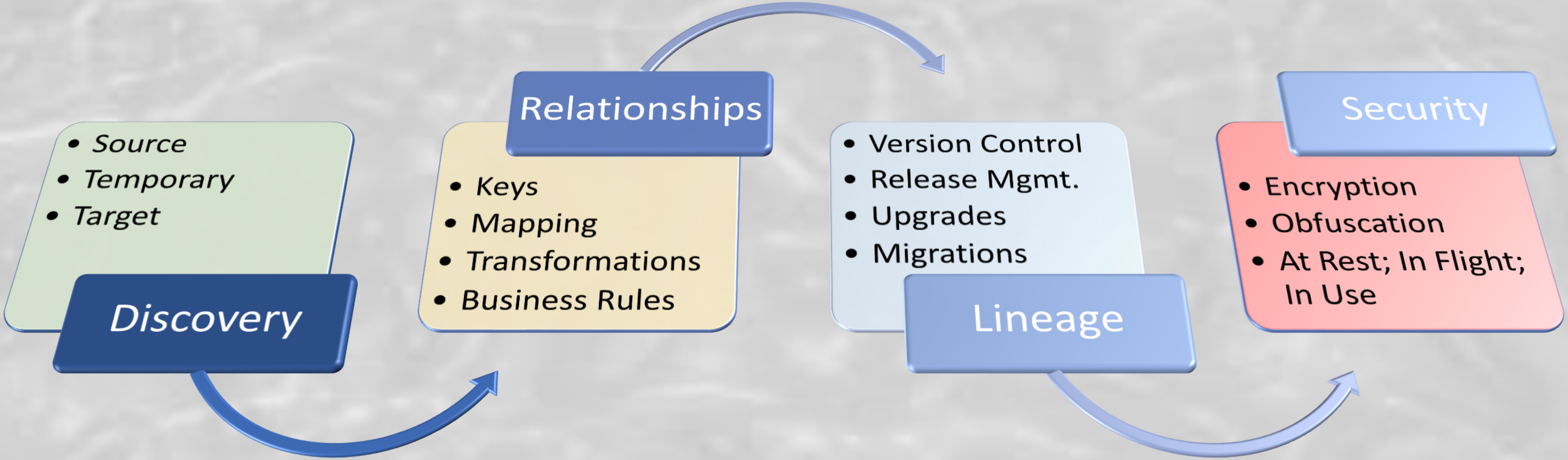
Typically, data accepted and processed by Business Rules; *whenever automation is possible*

Data PROFILING – Automated & Manual

Understand the Data; *size & shape + 3 V's*

Metadata Management

"Harvesting and Maintaining Metadata about Enterprise Systems data"



Discovery
- *Source*
- *Temporary*
- *Target*

Relationships
- Keys
- Mapping
- Transformations
- Business Rules

Lineage
- Version Control
- Release Mgmt.
- Upgrades
- Migrations

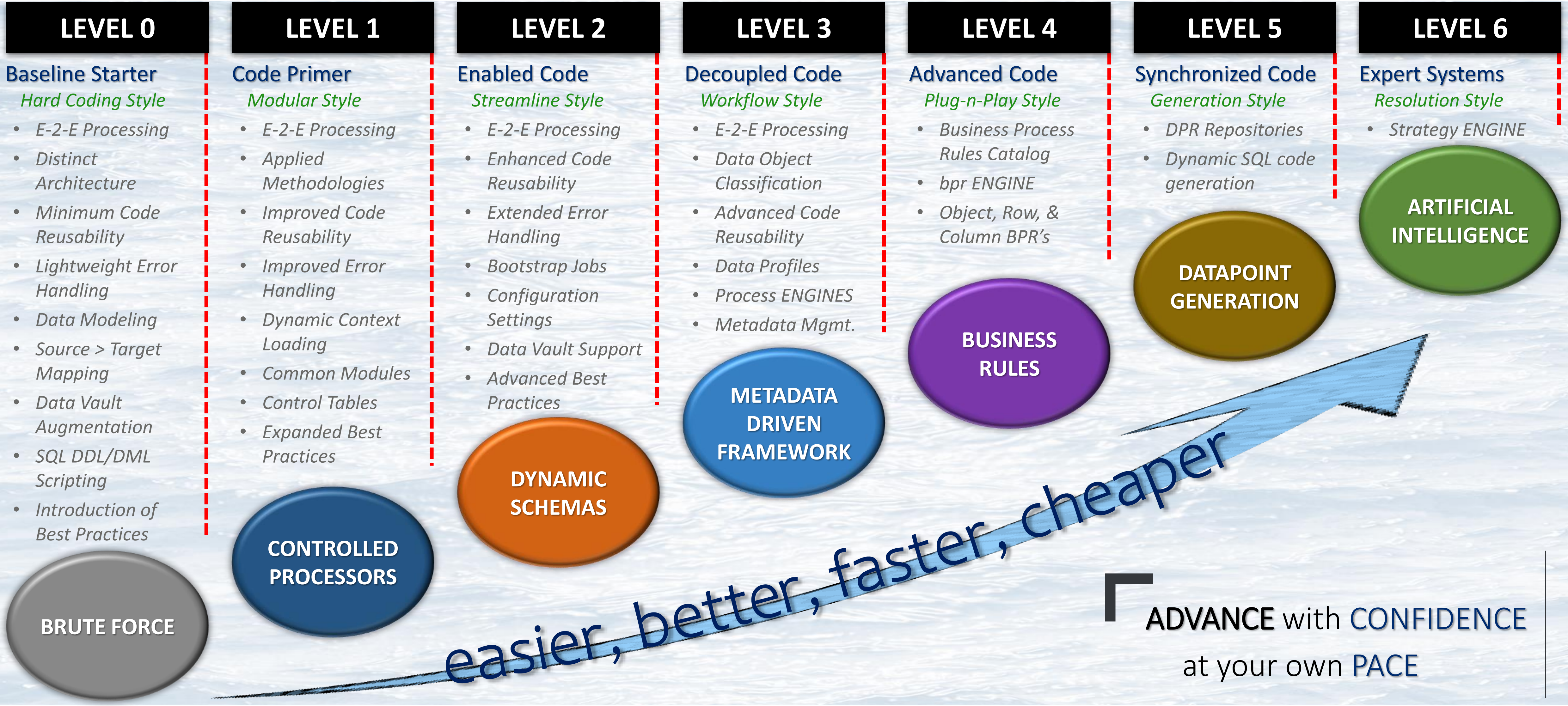
Security
- Encryption
- Obfuscation
- At Rest; In Flight; In Use

WORKFLOW AUTOMATION

- *Automation Maturity Model*
- *Modular Code Techniques*
- *Processor Control*

Automation Maturity Model

SCALEFREE
DATA DREAMLAND 2022



LEVEL 0

Baseline Starter

Hard Coding Style

- *E-2-E Processing*
- *Distinct Architecture*
- *Minimum Code Reusability*
- *Lightweight Error Handling*
- *Data Modeling*
- *Source > Target Mapping*
- *Data Vault Augmentation*
- *SQL DDL/DML Scripting*
- *Introduction of Best Practices*

BRUTE FORCE

LEVEL 1

Code Primer

Modular Style

- *E-2-E Processing*
- *Applied Methodologies*
- *Improved Code Reusability*
- *Improved Error Handling*
- *Dynamic Context Loading*
- *Common Modules*
- *Control Tables*
- *Expanded Best Practices*

CONTROLLED PROCESSORS

LEVEL 2

Enabled Code

Streamline Style

- *E-2-E Processing*
- *Enhanced Code Reusability*
- *Extended Error Handling*
- *Bootstrap Jobs*
- *Configuration Settings*
- *Data Vault Support*
- *Advanced Best Practices*

DYNAMIC SCHEMAS

LEVEL 3

Decoupled Code

Workflow Style

- *E-2-E Processing*
- *Data Object Classification*
- *Advanced Code Reusability*
- *Data Profiles*
- *Process ENGINES*
- *Metadata Mgmt.*

METADATA DRIVEN FRAMEWORK

LEVEL 4

Advanced Code

Plug-n-Play Style

- *Business Process Rules Catalog*
- *bpr ENGINE*
- *Object, Row, & Column BPR's*

BUSINESS RULES

LEVEL 5

Synchronized Code

Generation Style

- *DPR Repositories*
- *Dynamic SQL code generation*

DATAPOINT GENERATION

LEVEL 6

Expert Systems

Resolution Style

- *Strategy ENGINE*

ARTIFICIAL INTELLIGENCE

ADVANCE with CONFIDENCE

at your own PACE

Modular Code Techniques

Talend Reusable Jobs & Joblets

Encapsulate specific logic as atomic workflow components; *creates orchestration and reusable code modules that can be shared within and across PROJETS and JOBs*

Java Code Routines

Basic to highly complex methods that streamline expression handling and/or data transformation workflows; *creates simplified functions, incorporated where needed*

Snowflake UDF's

Enhances READ/WRITE functionality for dynamic SQL; *adds value to ELT data transformations*

External Executables/Scripts

Command Line processing capabilities; *establishes the ability to execute any independent workflow process*

Processor Control

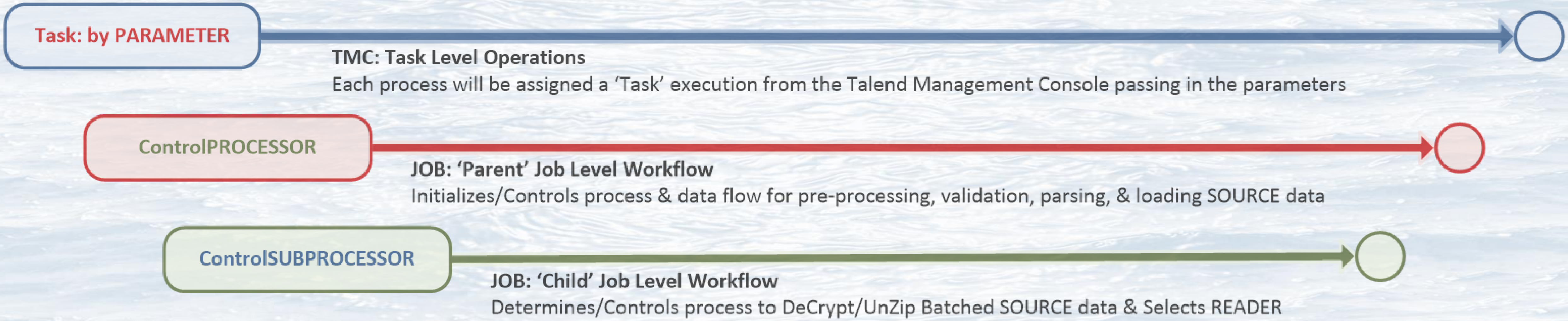
LEVEL 1 Automation Control Processors

TASK Management

JOB Processing; *establishes the task execution schedule & runs the Control Processor*

Control PROCESSORS

BATCH Processing; establishes the controlled batch workflow of SOURCE data to process



Task: by PARAMETER

TMC: Task Level Operations
Each process will be assigned a 'Task' execution from the Talend Management Console passing in the parameters

ControlPROCESSOR

JOB: 'Parent' Job Level Workflow
Initializes/Controls process & data flow for pre-processing, validation, parsing, & loading SOURCE data

ControlSUBPROCESSOR

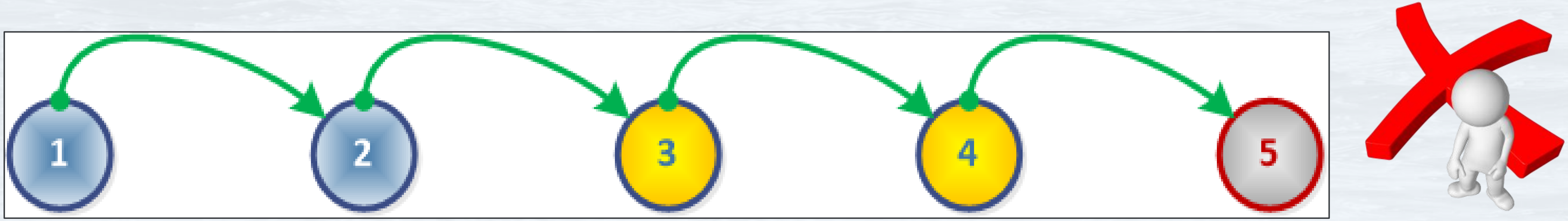
JOB: 'Child' Job Level Workflow
Determines/Controls process to DeCrypt/UnZip Batched SOURCE data & Selects READER

PATTERNS & PRACTICES

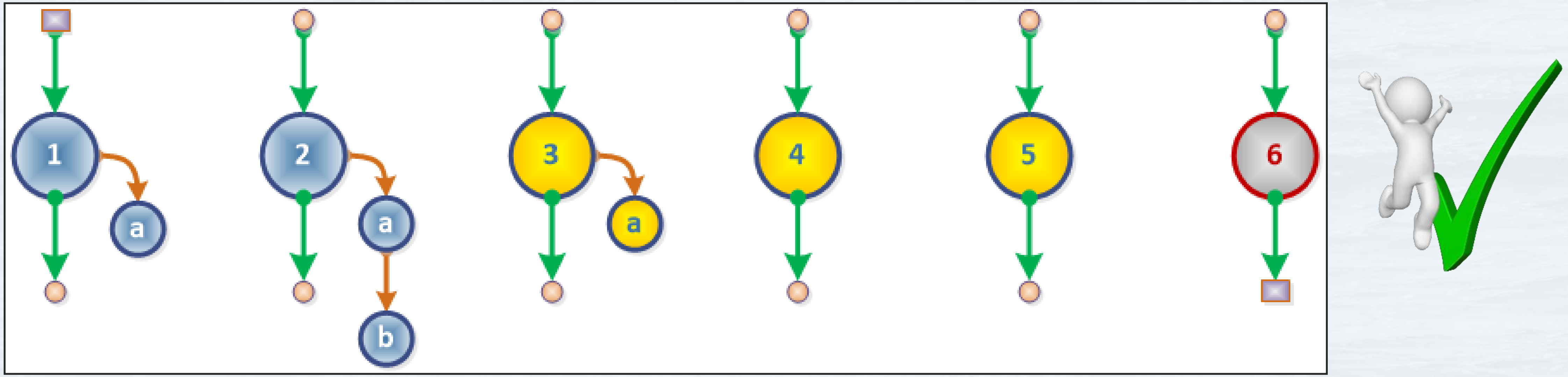
- *Job Design Patterns*
- *Process & Data Workflows*

Job Design Patterns

Coupled Process/Data flow: *creates dependencies; hard to maintain*



Decoupled Process/Data flow: *creates flexibilities; adaptable to change*



Process & Data Workflows

LEVEL 1 Automation
Get/Move JSON to TLZ/PSA

GET JSON Header

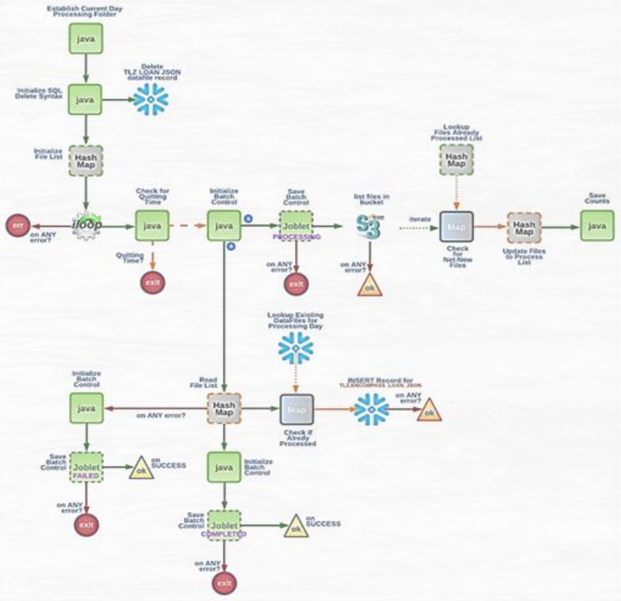
Locates JSON datafiles & creates TLZ temporary records

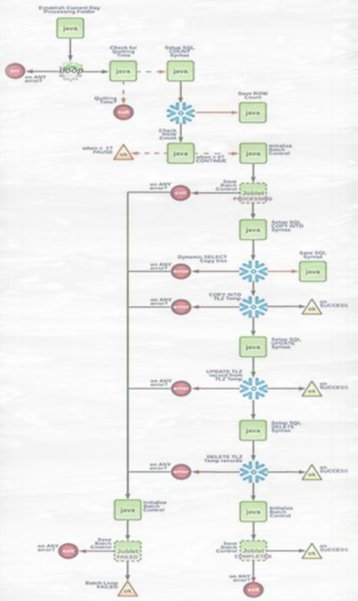
GET JSON Body

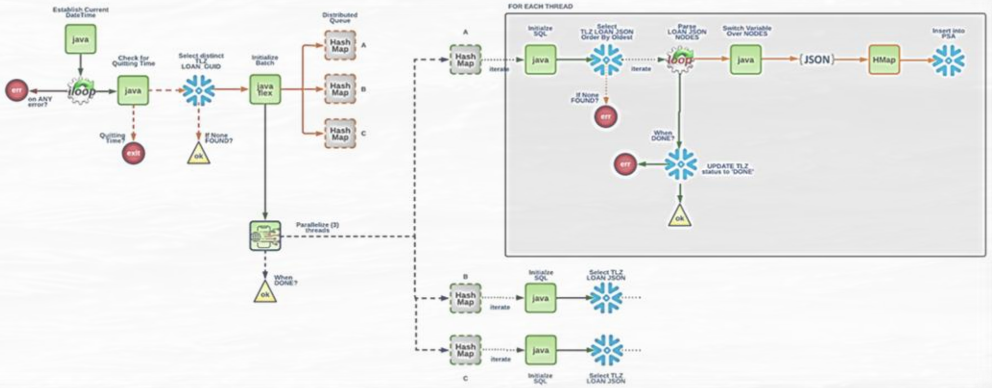
Adds JSON Body to TLZ temporary records

MOVE to PSA

Parses TLZ JSON records to PSA storage for downstream processing







Control PROCESSORS

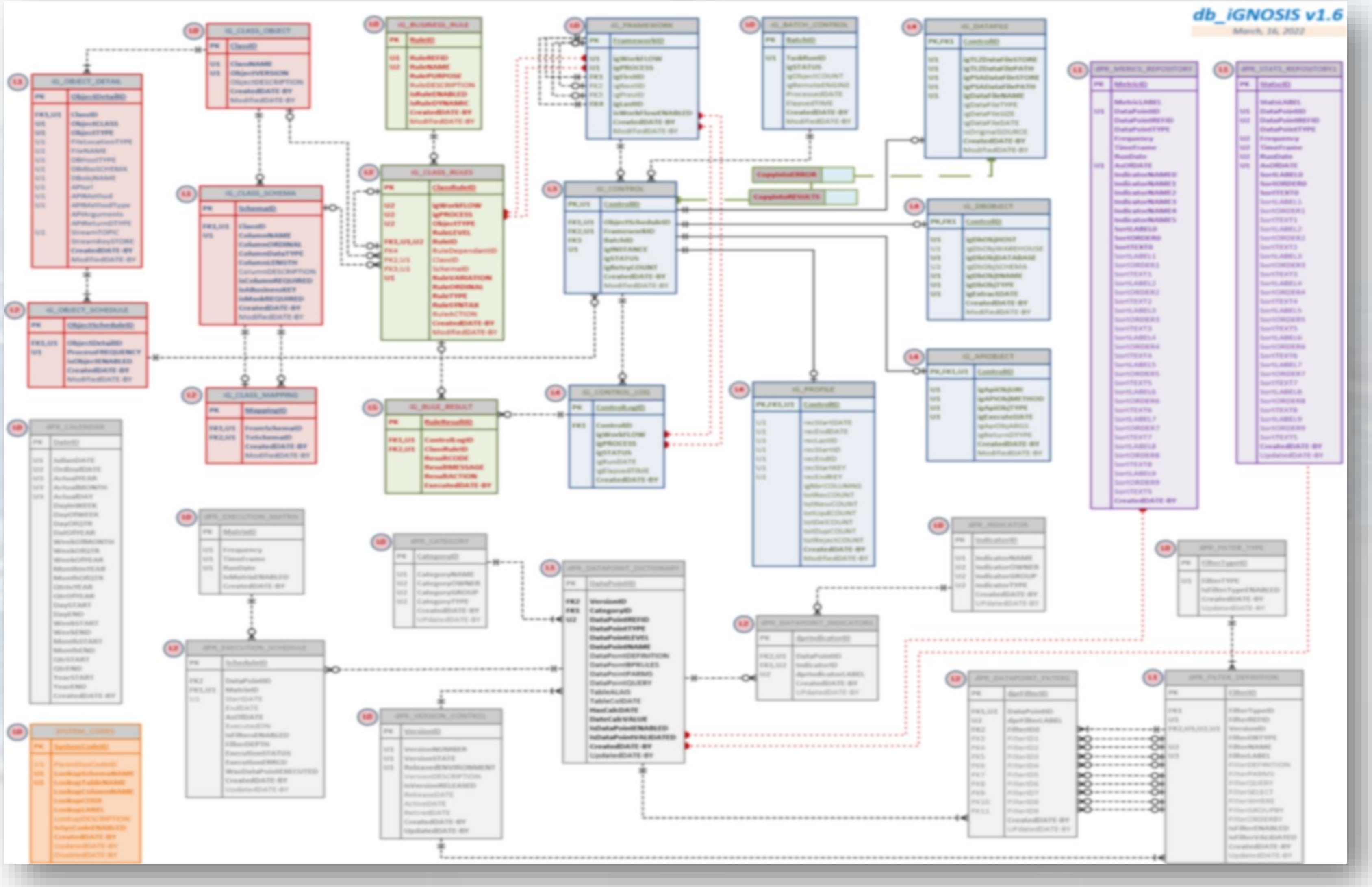
Controlled JOB Workflow; each JOB is designed to process INCREMENTAL or FULL data loads

Scalable Processing; each JOB is designed to run independently

METADATA DRIVEN FRAMEWORKS

- *Metadata Schema*
- *Automation Workflow Processors*

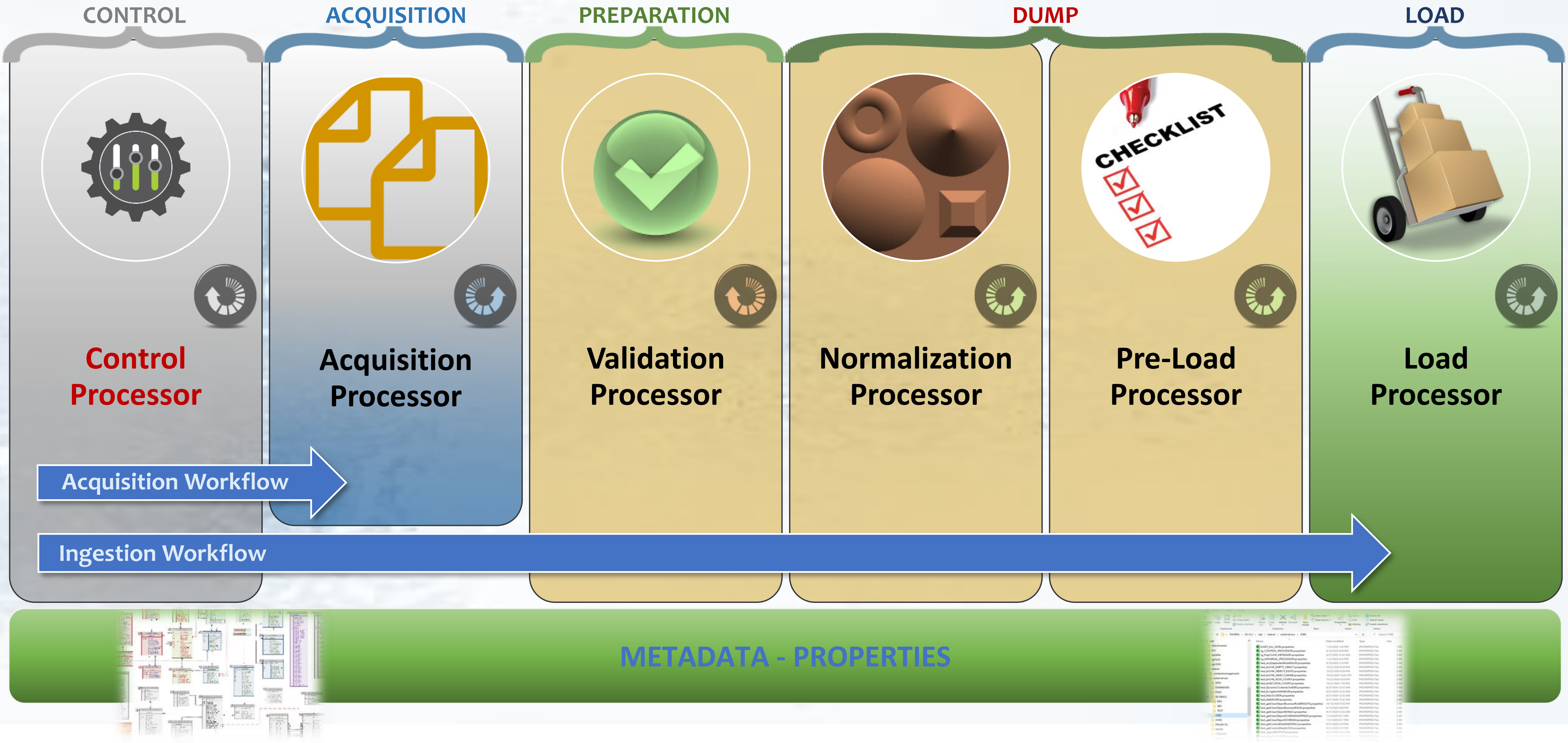
Metadata Schema



- CLASS OBJECTS [RED]
  - *Classification of any Data Object*
  - *Versioned Data Object Schema*
  - *Data Object Characteristics*
  - *SOURCE>TARGET Mappings*
  - *Process Schedule*
- BUSINESS PROCESS RULES [GREEN]
  - *Registration of any Business Rule*
  - *Assignment to Object/Result Set/Item*
  - *Records Execution Results*
- WORKFLOW CONTROLS [BLUE]
  - *Defines Framework Workflow Steps*
  - *Records Class Object Workflow State*
  - *Captures each Closs Object Details*
  - *Records Process Results*
- DATAPOINT REPOSITORY [GREY]
  - *Registration of any Datapoint*
  - *Datapoint Categorization*
  - *Execution Matrix*
  - *Filter Assignments*
  - *Operational Schedule*

Automation Workflow Processors

SCALEFREE
DATA DREAMLAND 2022



CONTROL

ACQUISITION

PREPARATION

DUMP

LOAD

Control Processor

Acquisition Processor

Validation Processor

Normalization Processor

Pre-Load Processor

Load Processor

Acquisition Workflow

Ingestion Workflow

METADATA - PROPERTIES

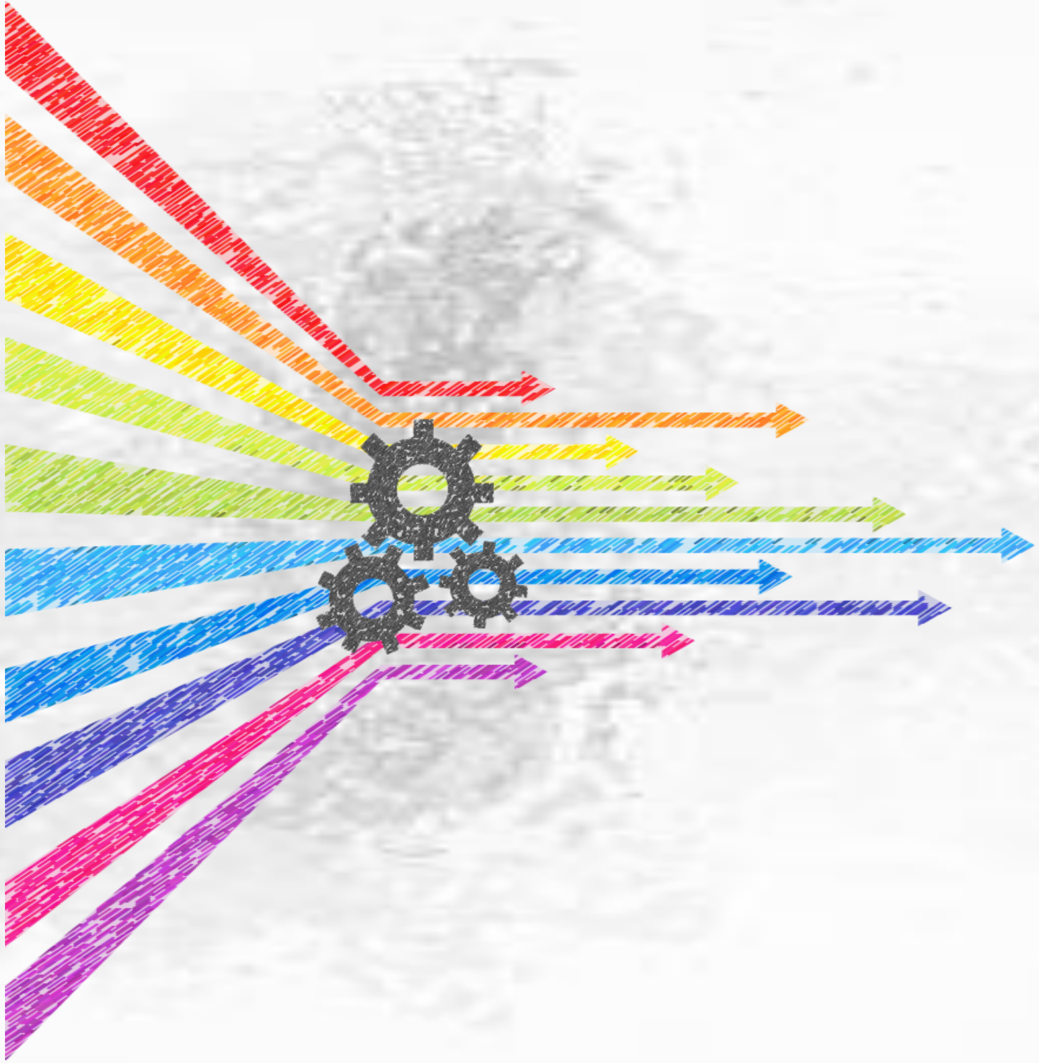
PLUG-N-PLAY BUSINESS RULES

- *Business Rules Catalog*
- *Business Process Rule Engine*

Business Rules Catalog

- Metadata Registration of BUSINESS RULES
  - ***class object level*** *can ACCEPT/REJECT object(s)*
  - ***schema row level*** *can ACCEPT/REJECT record(s)*
  - ***schema column*** *level can TRANSFORM value(s)*
  - ***schema cell level*** *can TRANSFORM a value*
- Rule TRANSFORMATIONS
  - *data goes in > transforms as coded > data comes out*
  - *data can be a value, a row, a column (cached), or a dataset*

Business Process Rule Engine



- Embedded in the UNIVERSAL Processor
  - *retrieves list of Business Process Rules (BPR) from metadata*
  - *controls execution of the BPR Engine iteratively*
  - *rule actions determine what happens after execution*
- TEST Rules
  - *typically designed to execute as a single business rule*
  - *an ordered list of business rules in sequence*
- COMMON Rules
  - *widely reusable, configurable with properties*
- OPERATIONAL Rules
  - *often designed to read/write data*
  - *others designed to operation at the class object level*

DATAPOINT REPOSITORY

- *Organized Datapoints*
- *Execution Matrix*
- *Information Marts*

Organized Datapoints

- Datapoint Dictionary
  - *unique name & identifier*
  - *atomic aggregation: Metric / KPI / STAT*
  - *defines clear language of datapoint purpose*
  - *defines SQL query template to produce desired values*
- Datapoint Categories
  - *unique name & identifier*
  - *specific grouping for datapoints*
- Datapoint Indicators
  - *unique name & identifier*
  - *vectored cross-correlations for datapoints*
    - ENGAGEMENT
    - GROWTH
    - PERFORMANCE
    - PROFITABILITY
    - RETENTION
    - UTILIZATION

Execution Matrix

"Specifies the reporting period for calculation of a datapoint"

Frequency

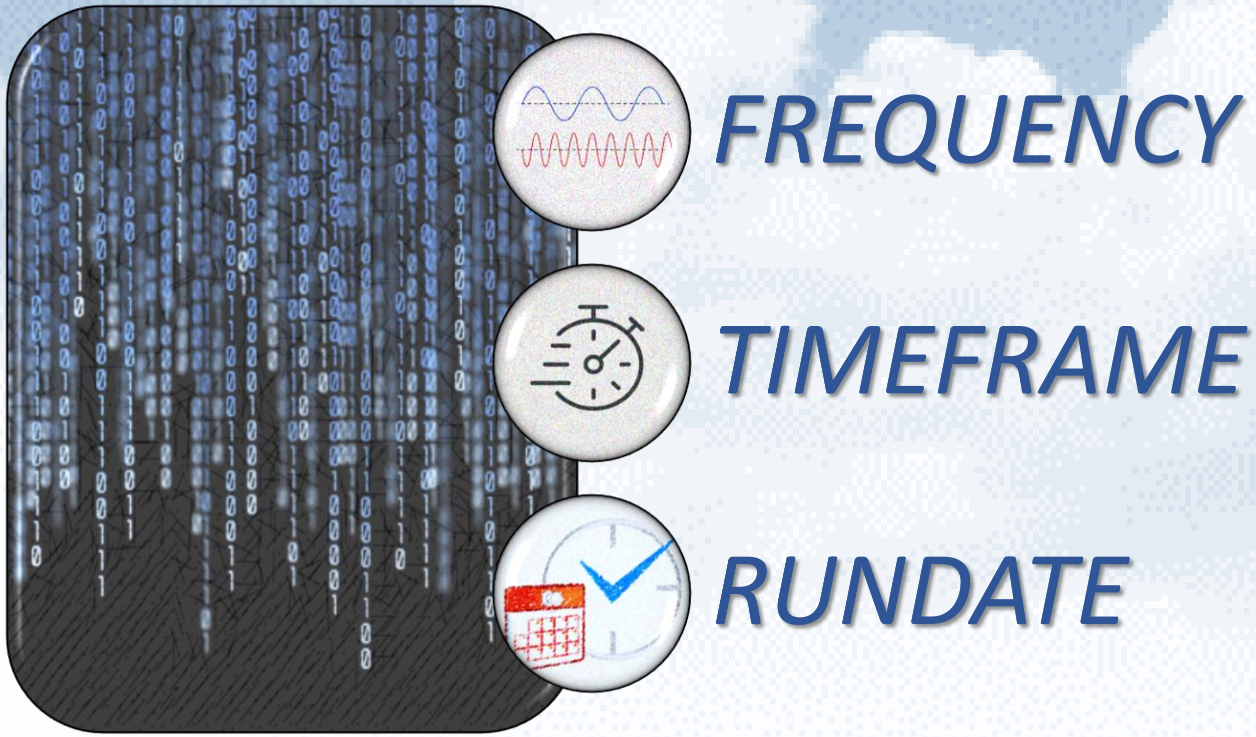
Specifies how often a datapoint is executed; hourly, daily, weekly, monthly, quarterly, annually

Timeframe

Specifies the date range for an executed datapoint; month-to-date, year-to-date, business-to-date

Run Date

Specifies the date/time when a datapoint is executed; beginning-of-period, end-of-period, in-period



FREQUENCY

TIMEFRAME

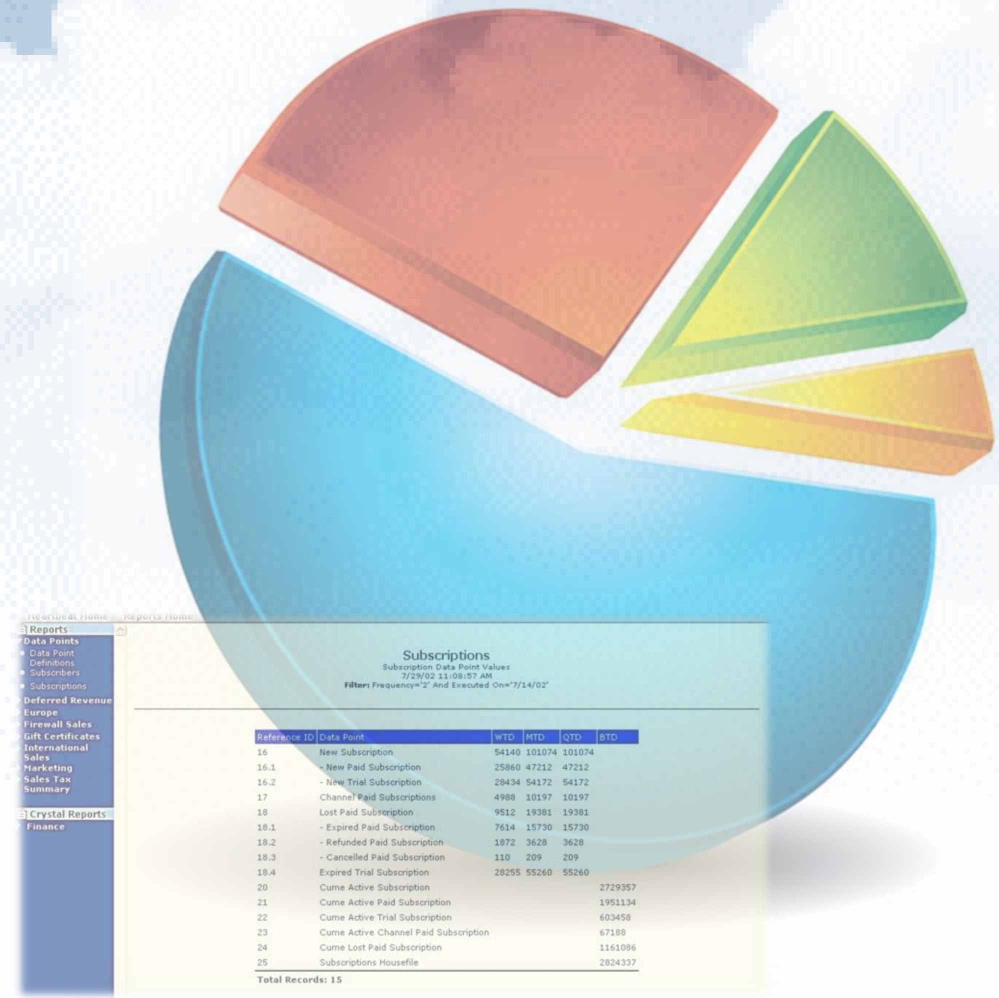
RUNDATE

FREQUENCY	TIMEFRAME	RUNDATE
Weekly Monthly Quarterly Annually	Week to Date Month to Date Quarter to Date Year to Date Balance of Month Balance of Year	Beginning of Period End of Period Prior to Period Within Period

Information Marts

“Pre-aggregated Datapoints streamline & simplify self-service reporting & interactive dashboards”

- Datapoints & Statistics
  - *dynamic SQL generation based upon defined template*
  - *each permutation result stored for:*
    - *execution matrix schedule*
    - *all defined filters for each datapoint*
  - *datapoint repository stores pre-aggregated results*
  - *datapoint dictionary & repository supports:*
    - *self-service reporting*
    - *interactive dashboards*
  - *multi-level datapoints delivers inherent quality checks*



Subscriptions

Subscription Data Point Values

7/29/02 11:08:57 AM

Filter: Frequency='2' And Executed On='7/14/02'

Reference ID	Data Point	WTD	MTD	QTD	BTD
16	New Subscription	54140	101074	101074	
16.1	- New Paid Subscription	25860	47212	47212	
16.2	- New Trial Subscription	28434	54172	54172	
17	Channel Paid Subscriptions	4988	10197	10197	
18	Lost Paid Subscription	9512	19381	19381	
18.1	- Expired Paid Subscription	7614	15730	15730	
18.2	- Refunded Paid Subscription	1872	3628	3628	
18.3	- Cancelled Paid Subscription	110	209	209	
18.4	Expired Trial Subscription	28255	55260	55260	
20	Cume Active Subscription				2729357
21	Cume Active Paid Subscription				1951134
22	Cume Active Trial Subscription				603458
23	Cume Active Channel Paid Subscription				67188
24	Cume Lost Paid Subscription				1161086
25	Subscriptions Housefile				2824337

Total Records: 15

SUMMARY

- *An **Agile Data Lake** delivers business value from disparate data sources: **easier**, **better**, **faster**, & **cheaper***
- *While **Data Governance** and **Automation** present challenges, when applied correctly they help deliver a trusted platform for all users*
- *Forging the **future of Data Engineering** projects require a collective of business and technical expertise*

QUESTIONS?

SCALEFREE

DATA
DREAMLAND
2022

Intelligent
Data Cloud

Dale Anderson
1+ (408) 375-8371
dale@trinityidc.com
www.intelligentdatacloud.com