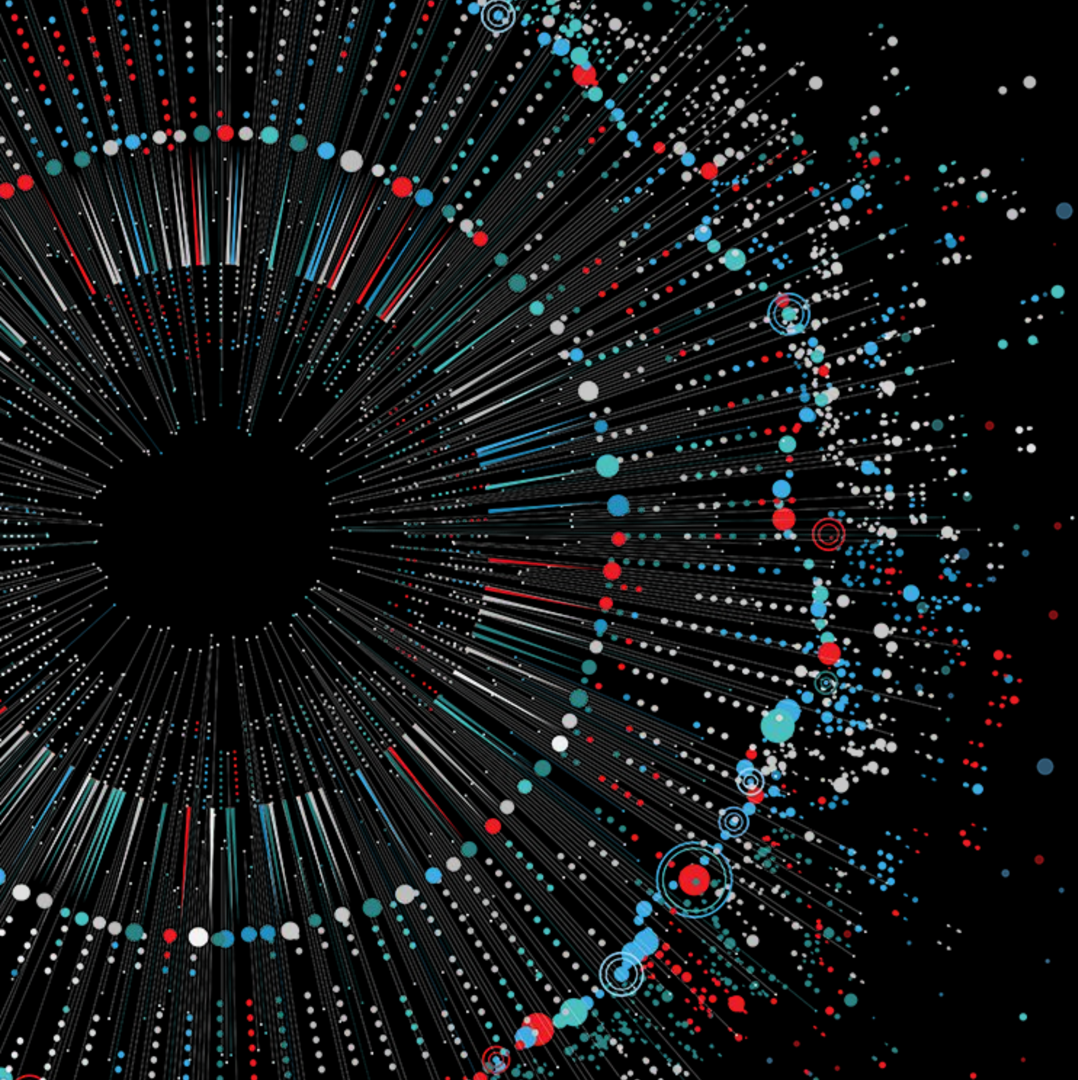


CRUX

Accelerate Source to Signal: Data Engineering Efficiency

Data Council NYC
Mark Etherington

November 2019 | Confidential

Drowning
in
Data

Evolving to an Efficient, Data-driven Operating Model is an Imperative but Very Challenging for Companies.

1. We've moved firmly into a **data driven world**. Companies that can get their hands on **more data, quickly and efficiently, have an edge.**

2. The "first mile" of **data ingestion is frustrating**. Managing data suppliers, extracting datasets from hundreds of sources, validating and wrangling data, and delivering to multiple destinations is **challenging and inefficient**.

3. The **vast majority of companies are on their own** and require a **robust technology platform** and **experienced operations team**.

4. Maintaining this data supply is a process that, if it fails is harmful to their business and, if it succeeds, **doesn't differentiate them from competitors.**

5. While mission critical, these capabilities are **expensive to build** and **detract resources** from differentiated work causing a **resource drain at some companies** and a **capability gap** at others.

6. **Margin pressure** has placed a new **focus on costs** across the financial services industry.

Major Data Trends

1. **Evolving to a data-driven model is critical but hard**

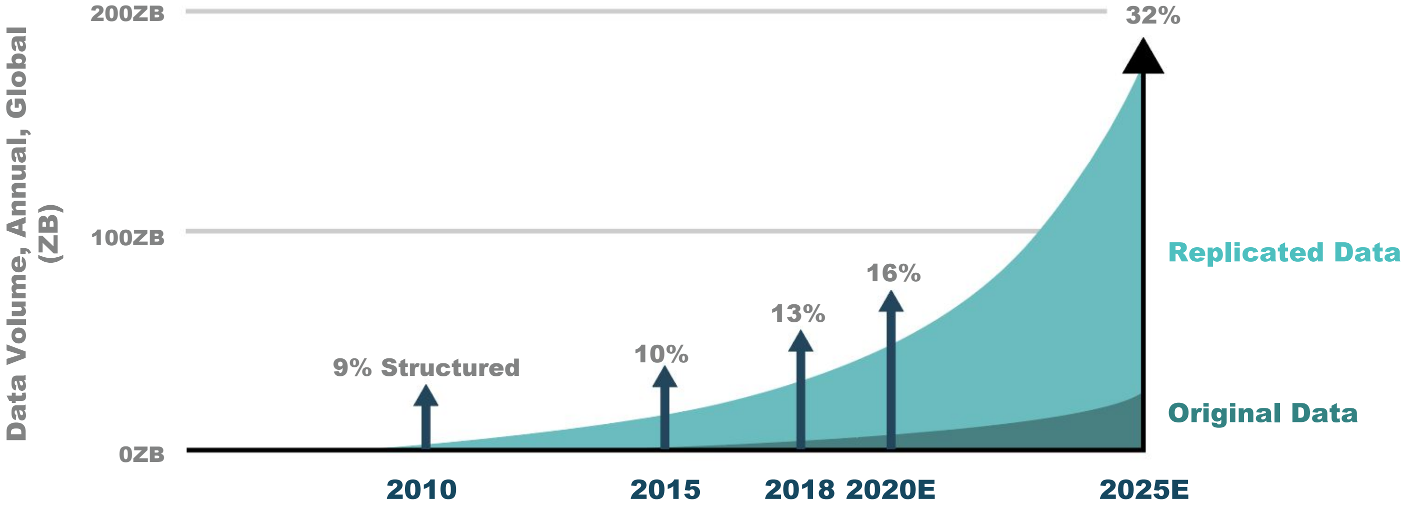
> **Companies Are Failing in Their Efforts to Become Data-Driven**
> *- Harvard Business Review, February 2019*

2. **Increasingly blurred lines b/w data creators and distributors**

> **London Stock Exchange clinches acquisition of Refinitiv for $27Bn**
> *- Financial Times, July 2019*

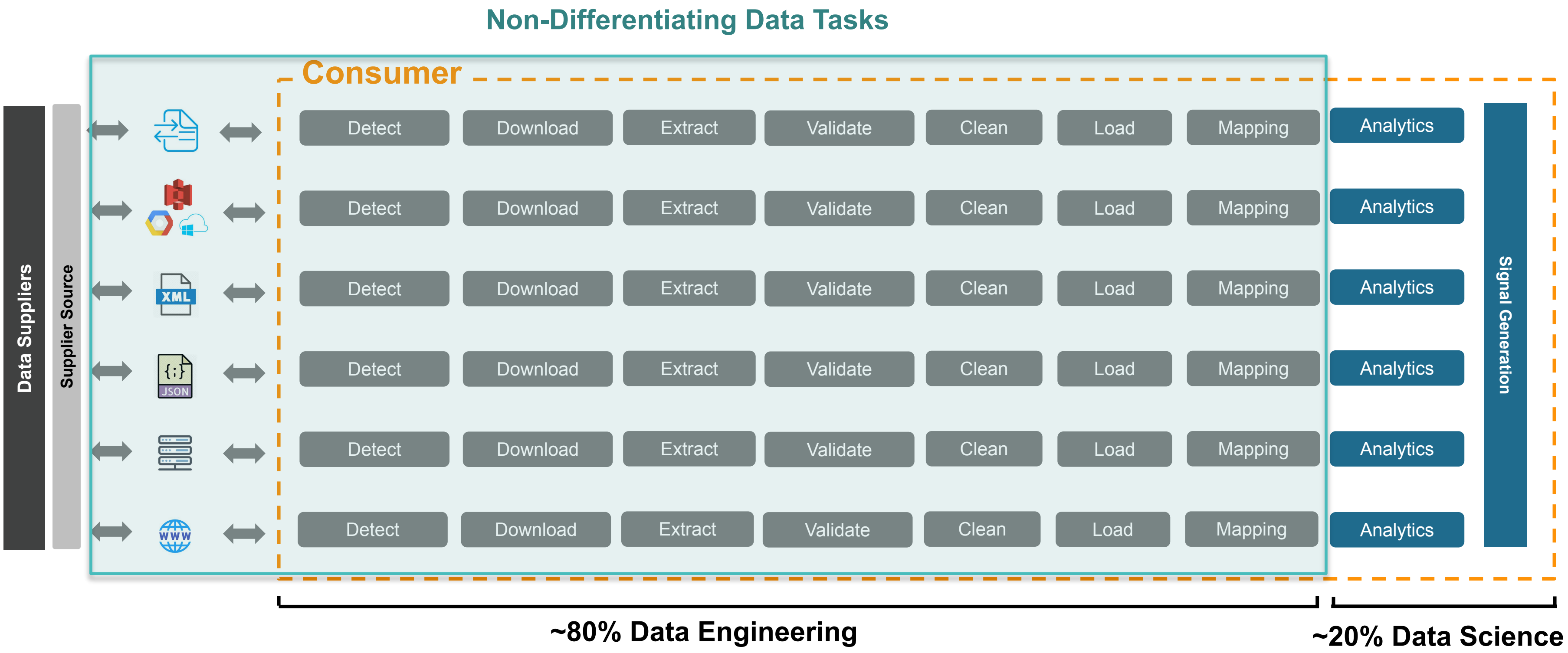
3. **Long path to monetization roadblock for new vendors.**

> **Hard times at Thasos might be sign of Alternative Data's Future**
> *- Business Insider, August 2019*



Source: IDC "Digitalization of the World From Edge to Core White Paper" developed in collaboration with Seagate (11/18), IDC DataSphere. Note: 1 petabyte = 1M gigabytes, 1 zetabyte = 1M petabytes of new data created/captured each year. The dark teal area in the graph represents data generated, not stored. Structured data indicates data that has been organized so that it is easily searchable and includes metadata & machine to machine (M2M) data. Replicated data = data that is a copy of the original.

Massive Firmwide Inefficiency



Non-Differentiating Data Tasks

Consumer

Data Suppliers

Supplier Source

XML

JSON

Detect	Download	Extract	Validate	Clean	Load	Mapping	Analytics
Detect	Download	Extract	Validate	Clean	Load	Mapping	Analytics
Detect	Download	Extract	Validate	Clean	Load	Mapping	Analytics
Detect	Download	Extract	Validate	Clean	Load	Mapping	Analytics
Detect	Download	Extract	Validate	Clean	Load	Mapping	Analytics
Detect	Download	Extract	Validate	Clean	Load	Mapping	Analytics

Signal Generation

~80% Data Engineering

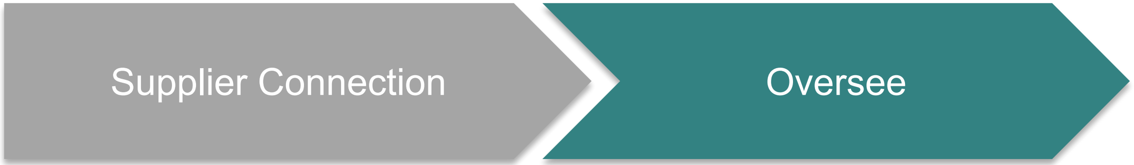
~20% Data Science

Pipelines Aren’t Just A Technical Problem

Supplier Connection

- Contracts & Relationships
- Fast Ingestion
- Any Data Type & Source
- Ready-made Pipelines

A ~~Puppy~~ Pipeline Isn’t Just For Christmas



Supplier Connection

Oversee

- Contracts & Relationships
- Fast Ingestion
- Any Data Type & Source
- Ready-made Pipelines

- **Data Validation**
- **Reliable Operations**
- **24/7 Support Desk**

Delivery Is Not a Given

Supplier Connection

- Contracts & Relationships
- Fast Ingestion
- Any Data Type & Source
- Ready-made Pipelines

Oversee

- Data Validation
- Reliable Operations
- 24/7 Support Desk

Deliver

- **Flexible Delivery Mechanism**
- **Entitlements & Analytics**
- **Notifications**

Client’s Are Different

Pull

- API
- Python client
- FTP

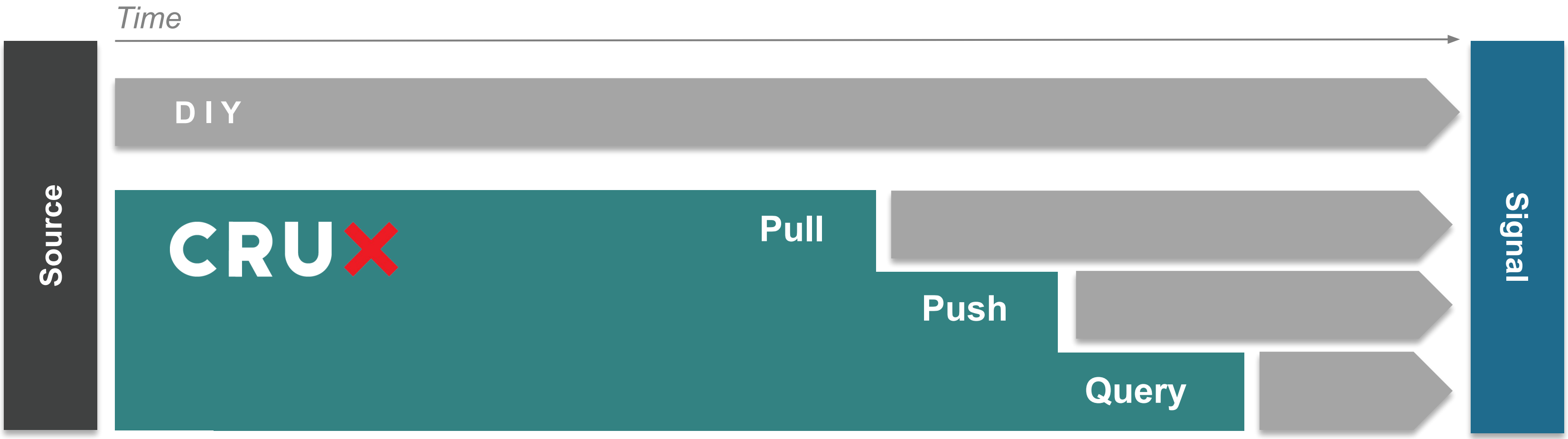
Push

- Client data lake
- Cloud warehouses (integrated with AWS, Snowflake, GCP)
- Analytics platforms

Query

- Query what you need

Client's Need to Focus on the Signal



THINK DIFFERENTLY

CRUX

The Ideal Solution

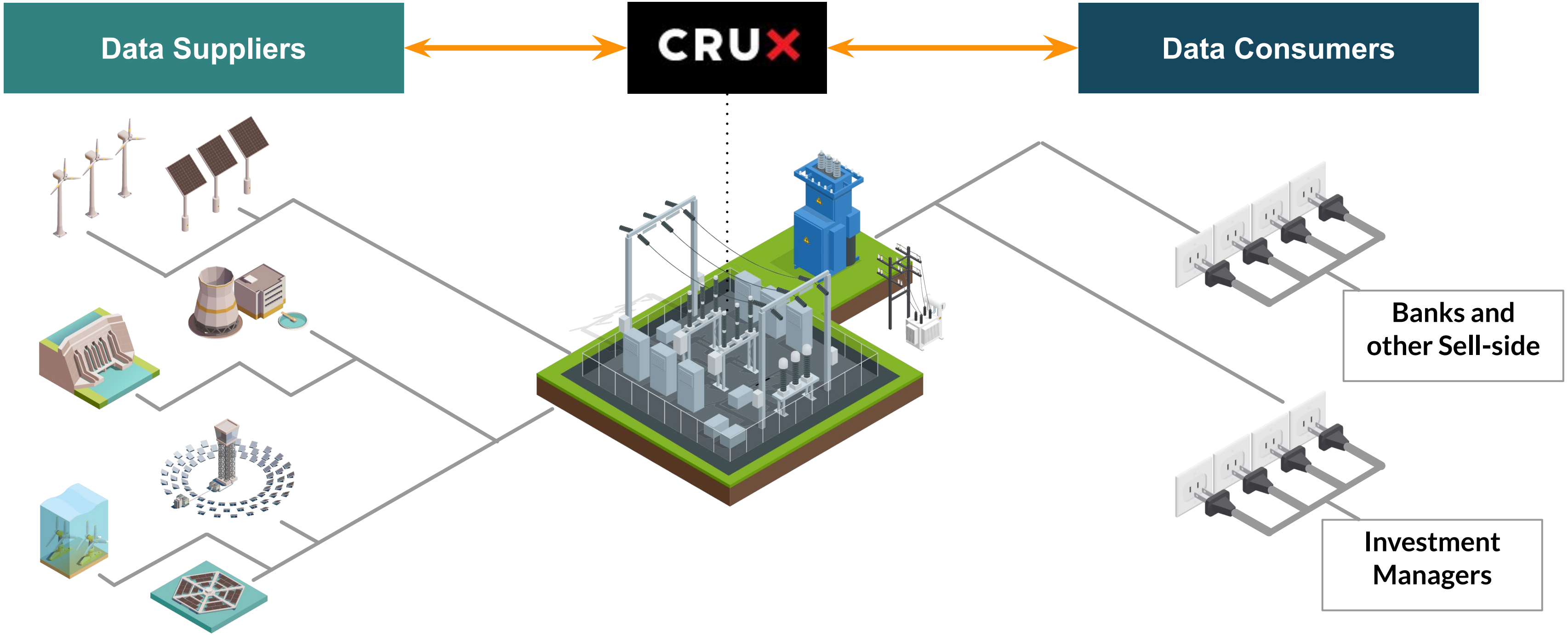
For Data Suppliers:

- Delivery for all formats to multiple destinations
- Technical expertise on data systems
- Neutrality from other data suppliers and brokers
- Meeting client expectations for their operations
- Reliable operations & support
- Future-proof technology
- Confidentiality & Security

For Data Consumers:

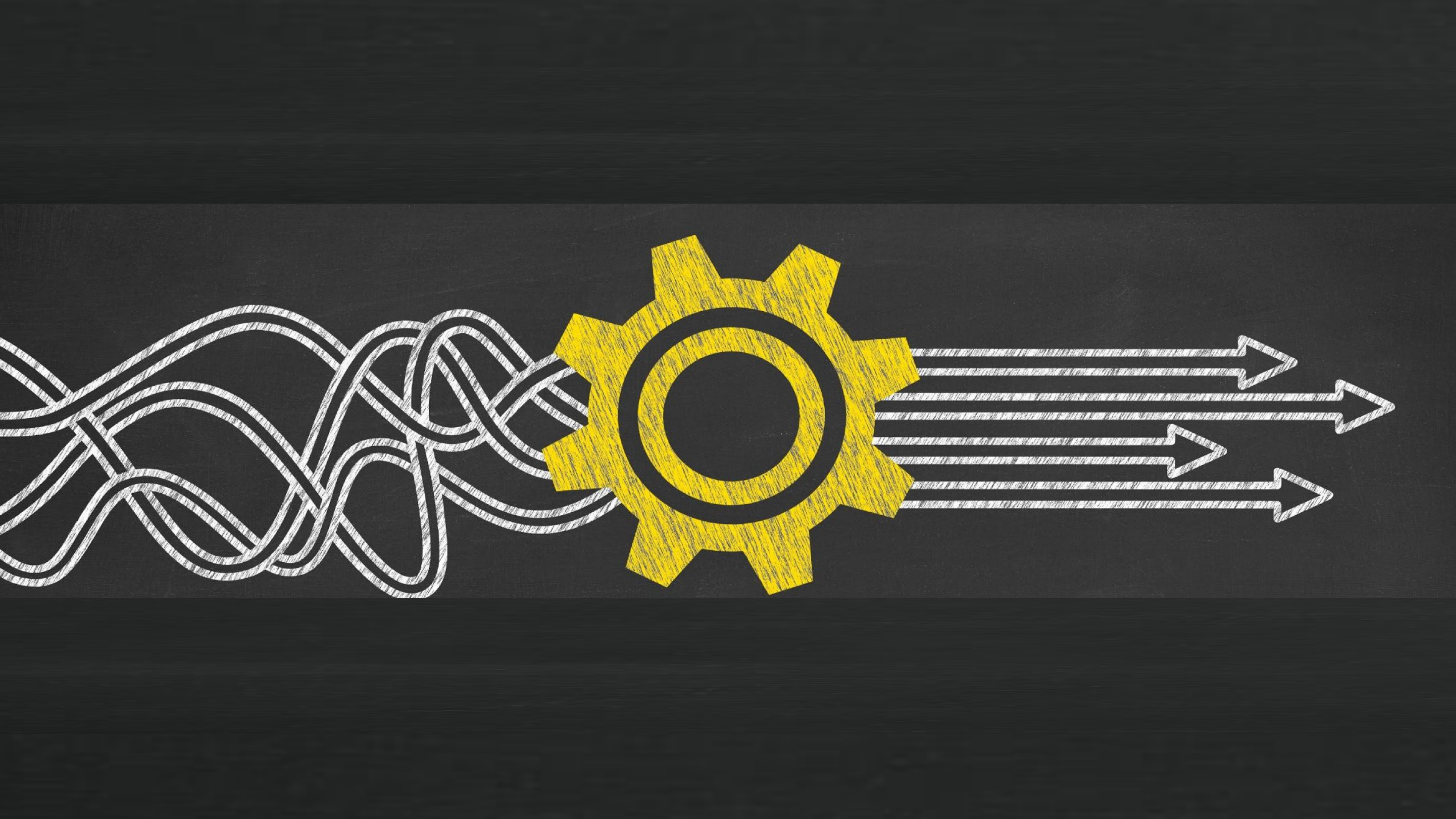
- Single access point to all data sources
- Technical expertise on datasets
- Scalable onboarding and access to data
- Delivery to desired team at desired endpoint
- Reliable operations & support
- Future-proof technology
- Confidentiality & Security

Economies Are Needed Across the Data Supply Chain



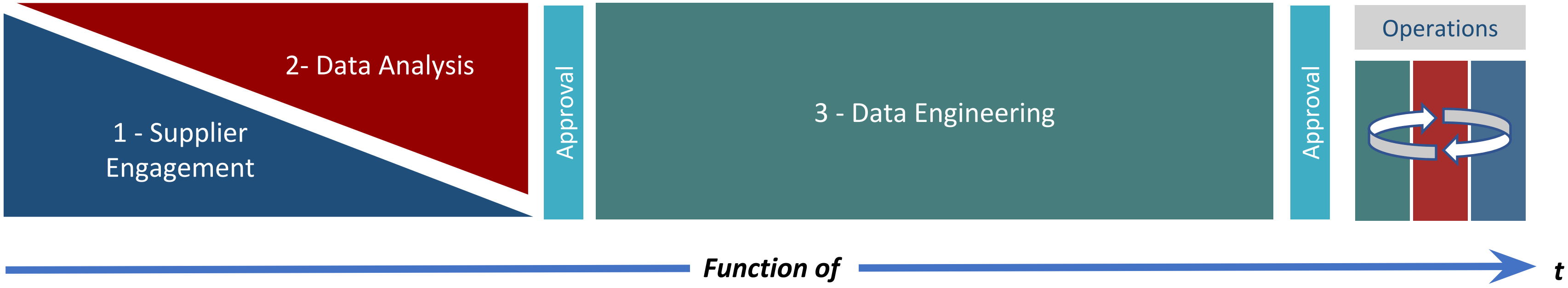
Crux has amassed the largest repository of supplier agreements across the entire data industry

As an industry utility, the value that Crux adds to each client is amplified across the entire client universe



Time & Effort

Acquisition & Definition ***Technical Onboarding and Testing*** ***Production***



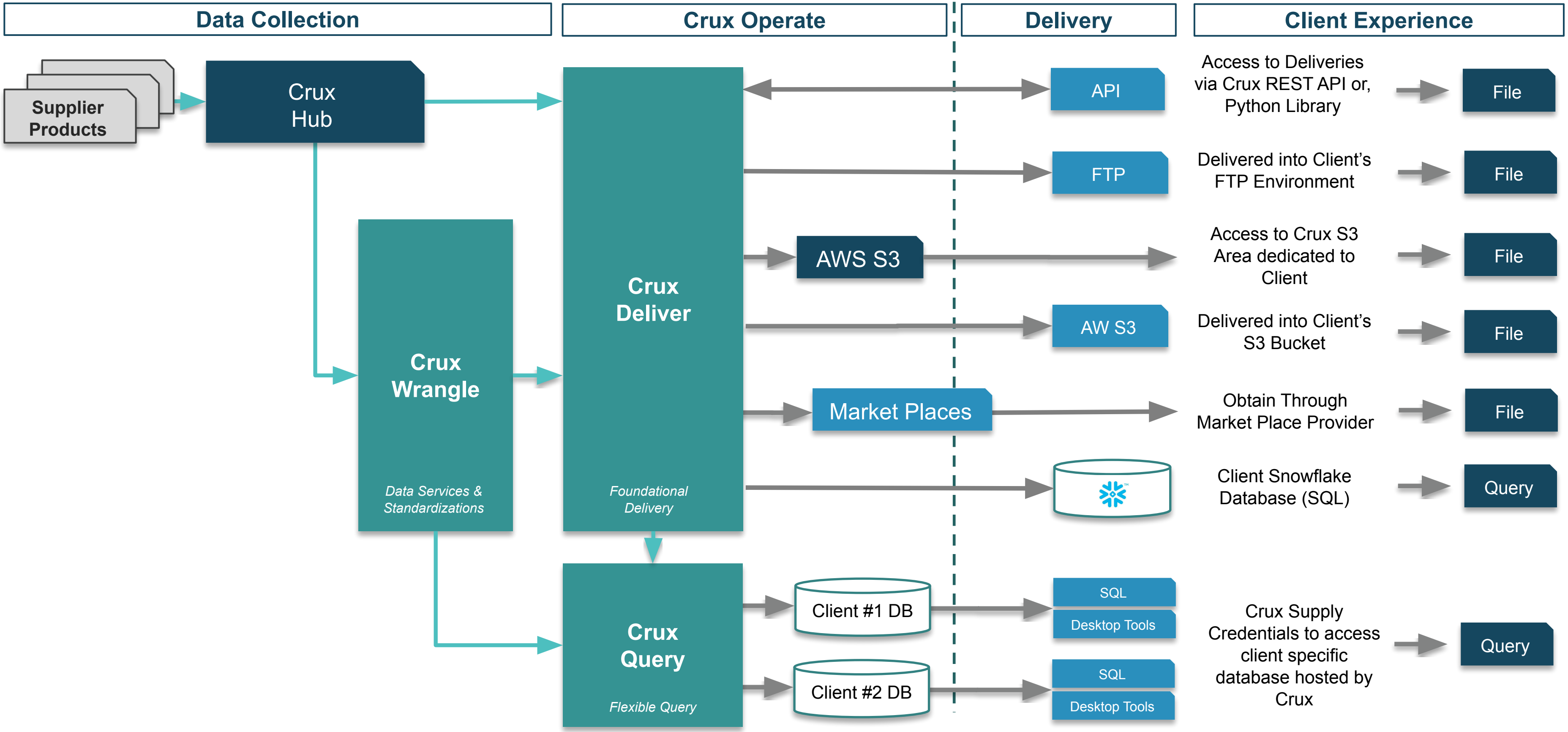
- Complexity & Size of “Dataset”
- Supplier Voracity
- Quality of Supplier Data & Infrastructure
- Client Participation

- Complexity & Size of “Dataset”
- Type of Source (File, Loader, API, Website)
- Encoding of Data (CSV, HTML, XLM, JSON)
- Quality of Definition

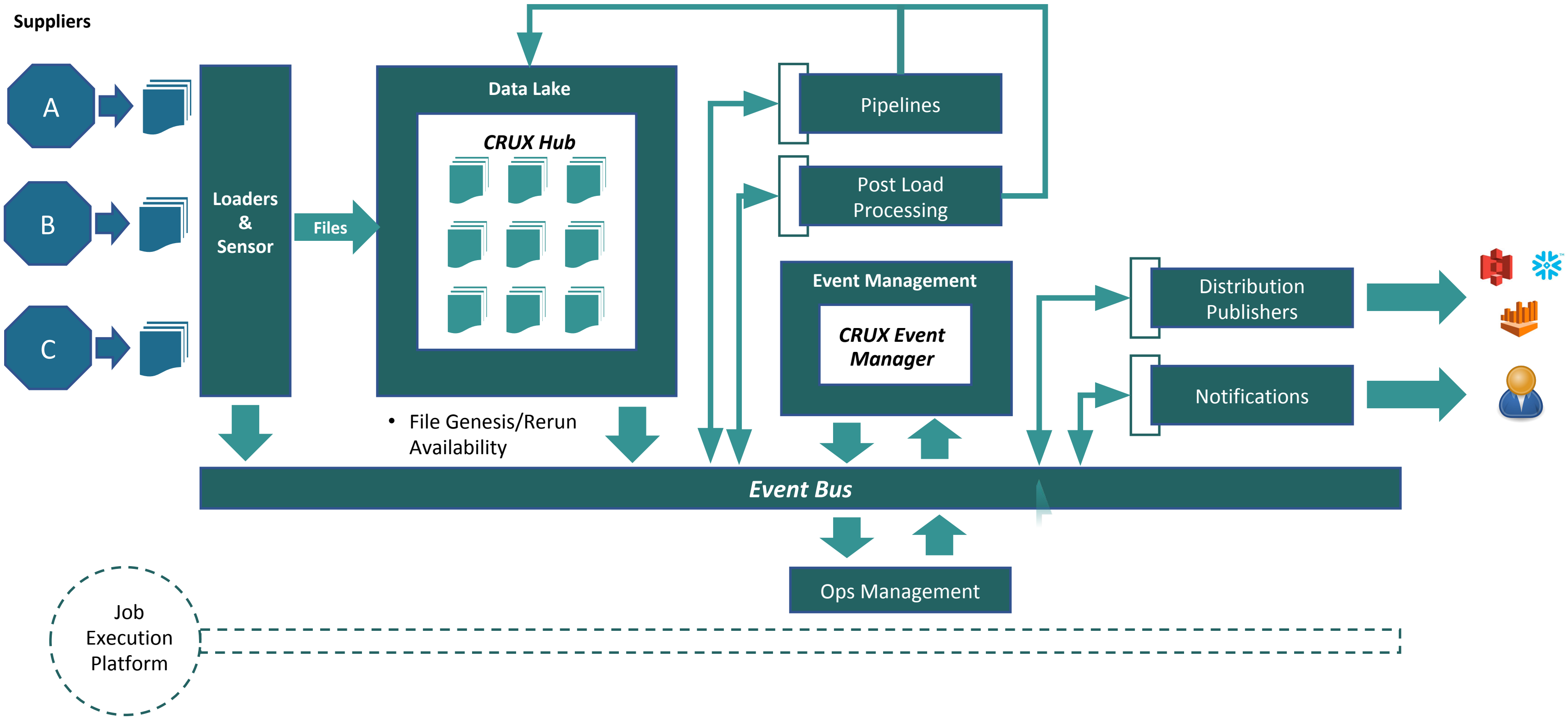
- Stability of:
  - Supplier Infra
  - Crux Infra
  - Data Schemas

- Resource Quantity
- Resource Quality
- Tooling
- Process Mgmt

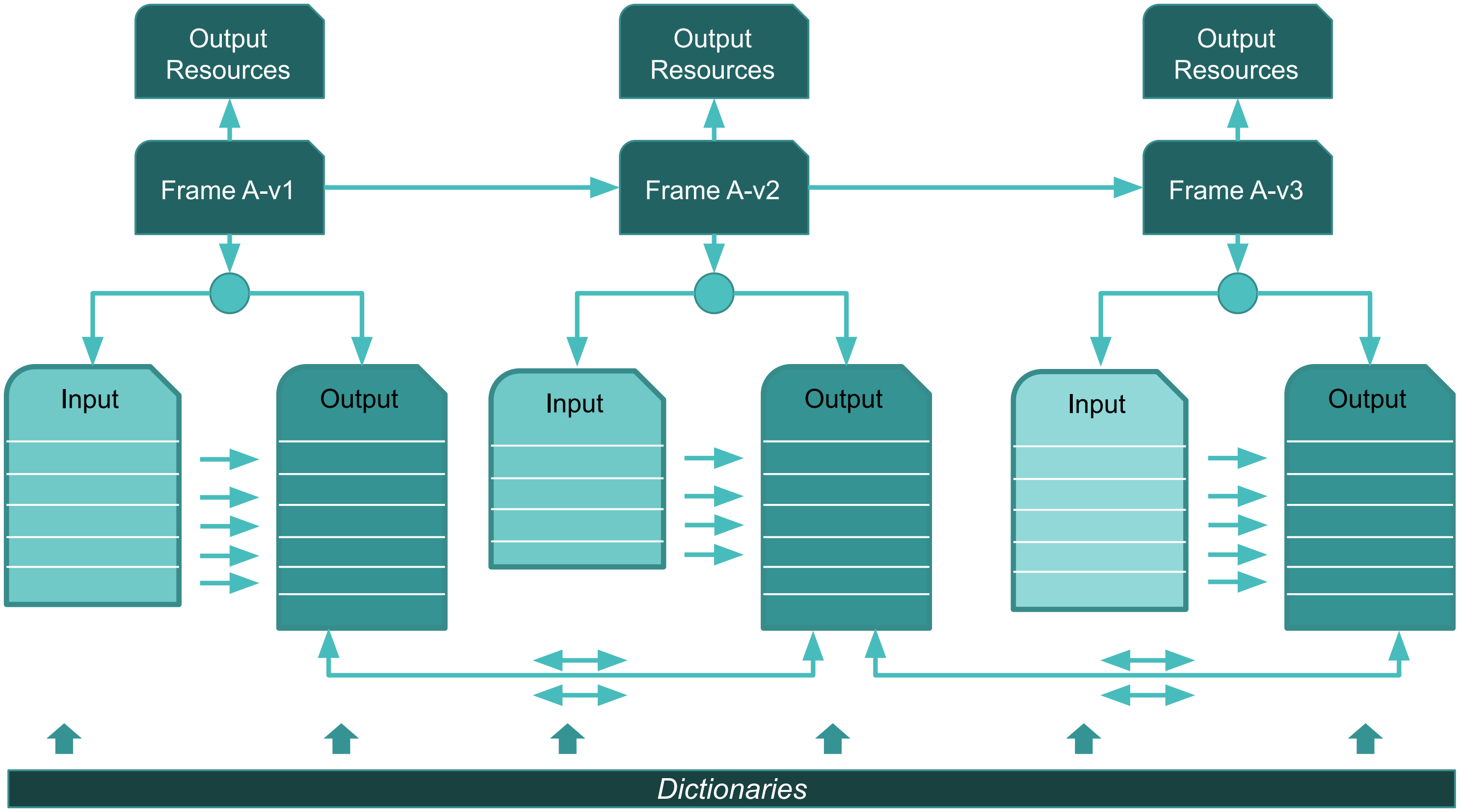
Crux Client Facing Products


Data Collection
Crux Operate
Delivery
Client Experience
Supplier Products
Crux Hub
Crux Wrangle
Data Services & Standardizations
Crux Deliver
Foundational Delivery
Crux Query
Flexible Query
API
Access to Deliveries via Crux REST API or, Python Library
File
FTP
Delivered into Client's FTP Environment
File
AWS S3
Access to Crux S3 Area dedicated to Client
File
AW S3
Delivered into Client's S3 Bucket
File
Market Places
Obtain Through Market Place Provider
File
Client Snowflake Database (SQL)
Query
Client #1 DB
Client #2 DB
SQL
Desktop Tools
SQL
Desktop Tools
Crux Supply Credentials to access client specific database hosted by Crux
Query

Event Architecture



Data Can Be Complex, Schemas are "Worse"



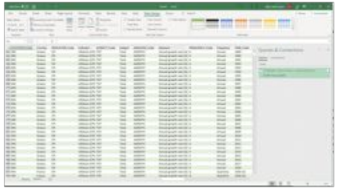
Key Insights – Data 101

- Data pipelines are the start, not the end, of the solution
- There’s no such thing as standard (if a behavior can happen, it will, twice!)
- Data insight is harder than data pipelining
- Supplier contracting can be harder than insight
- Distribution always has nuances

Key Insights – Technology 102

- Not everything is a “Big Data” Problem
- Not every solution needs a “Big Data” Tool
- Mature Technologies have their place
- Focus on own data – Understand your atoms
- Name Early, Bind Late
- Lies, Damn lies and Column names

Excel

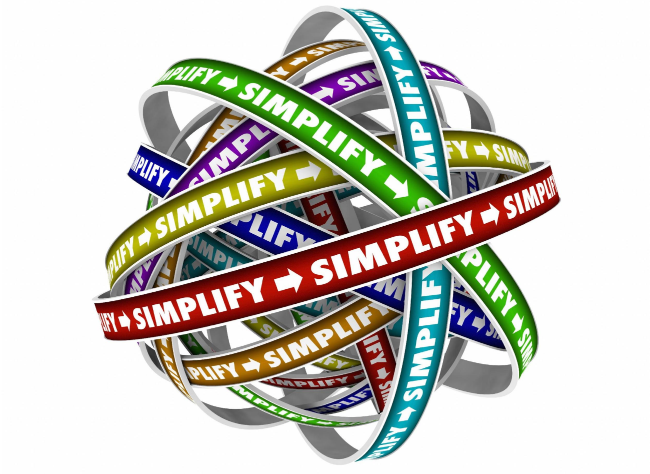


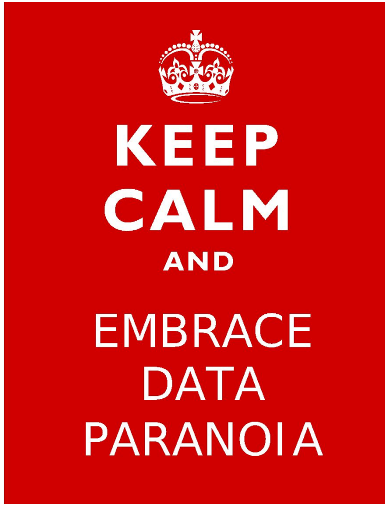
Tableau

Power BI

- Architecture cannot be brittle needs - roll with the punches
- Data duplication is not solving the client’s problem
- Dictionaries need to be Curated
- You can’t automate “stupidity”, but you can spot stupidity with automation

Guide to Data Karma





Thank You

data@cruxinformatics.com

Office Hours

3:15 - 4:00 PM

Room 476A