SIGMOID

Data Quality and Governance

www.sigmoid.com

About Sigmoid

Sigmoid is an emerging leader in data engineering and AI solutions.

750+
Employees

Work with **30+**
Fortune 500 firms

>97%
CSAT score

200+
ML models operationalized

5000+
Data pipelines built

Backed by
SEQUOIA

Awards and Recognition

500™ Technology Fast 500
2023 NORTH AMERICA
Deloitte.

ISG Provider Lens 2023 Quadrant
Analytics Services
Rising Star, U.S. ★
Report releasing Jan 2024

Inc. 5000 America's Fastest-Growing Private Companies

DATA BREAKTHROUGH AWARD 2021
Open Source data solution provider of the year

FORRESTER
Now Tech: AI Consultancies, Q1, 2021 Report

Major Contender in
EVEREST GROUP
Analytics and AI Services Specialists PEAK Matrix (2022)

Offices

New York

San Francisco

Dallas

Lima

Bangalore

Amsterdam

London

Sao Paulo

Enabling Business Transformation with Full-Service Capability Suite

Business Consulting & Data Strategy

- Data Strategy & Vision
- Data Monetization
- Data & Technology Roadmap
- Technology Evaluation & Selection
- Data Governance & Security Strategy
- AI/Gen AI Strategy

Data Engineering Services

Data Pipelines	ML Engineering	Cloud Trans.	BI / Consumption
Data Migration & Conversion	Model scaling & productionizing	Cloud Migration	Data Lake / Mesh
			Data Product
Performance Optimization	Feature Engineering	Application Modernization	BI Reporting & Visualization
Data Ingestion: ETL/ELT	Pipeline Optimization	Cost optimization	AI/ML, LLM

Data Science

- Supply Chain Analytics
- Marketing & Consumer Analytics
- Operational Analytics
- E-Commerce & Sales Analytics

Managed Services

- Data Labs
- Cloud Infra Support and Management
- Devops and Secops Support
- DataOps & ML Ops
- Data Application Managed Services

Governance & Security Services

- Data Catalog & Lineage
- Master Data Management
- Data Quality & Security

Technology Expertise

Technology Partners

Microsoft, databricks

Cloud Technologies

python, PySpark, jupyter, TensorFlow, R, NLP, MATILLION, DATADOG, Spark, hadoop, cassandra, mongoDB, APACHE HBASE, atlan, Alation, kubernetes

Capabilities on Azure Stack

Sigmoid has worked with large customers to design, build and deploy data products in Azure

Data Processing & Transformation:

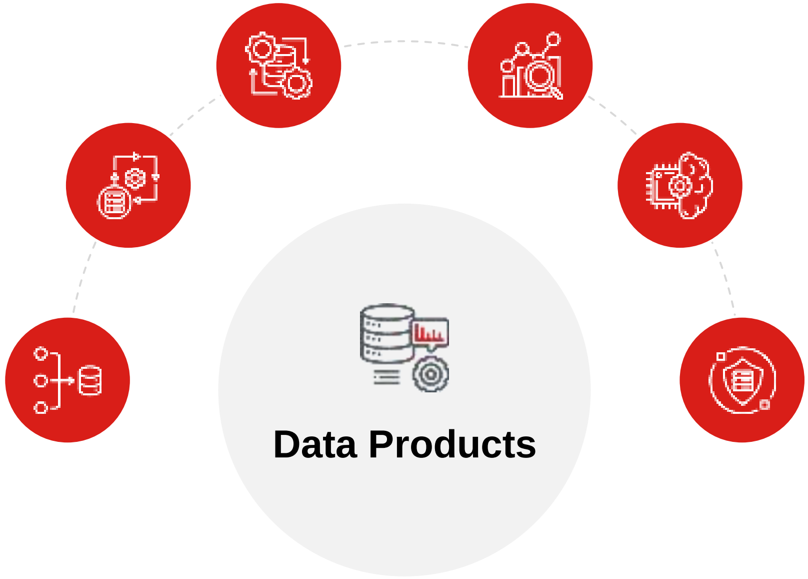
- **Azure Databricks:** Collaborative Apache Spark-based analytics platform to be used for big data processing and machine learning.
- **Azure HDInsight:** Managed cloud service for processing big data using popular open-source frameworks like Hadoop and Spark.

Data Storage & Management:

- **Azure Data Lake Storage:** Scalable and secure data lake for storing large amounts of structured and unstructured data would be considered.
- **Azure SQL Database:** Managed relational database service for structured data storage.

Data Ingestion & Integration:

- **Azure Data Factory:** Creating data pipelines to move and transform data from various sources.
- **Azure Event Hubs:** Real-time data ingestion from applications, devices, or any data streams would be done.



Data Products

Data Analytics & Visualization:

- **Azure Synapse Analytics:** Analytics service which will be used for analyzing large amounts of data using either serverless or provisioned resources.
- **Power BI:** Business intelligence tool to be used for creating interactive visualizations and reports.

Machine Learning & AI:

- **Azure Machine Learning:** End-to-end platform for building, training, and deploying machine learning models.
- **Cognitive Services:** Pre-built AI services for vision, speech, language, and decision-making.

Security & Compliance:

- **Azure Active Directory:** Identity and access management service.
- **Azure Policy & Blueprints:** Tools for implementing governance and compliance across Azure resources.

Sigmoid’s implementation of data products in Azure involves leveraging a combination of services and tools tailored to specific business needs. Sigmoid would collaborate between data engineers, data scientists, business analysts, and other stakeholders as it is essential to align the implementation with business goals and ensure success.

Principles of Data Governance

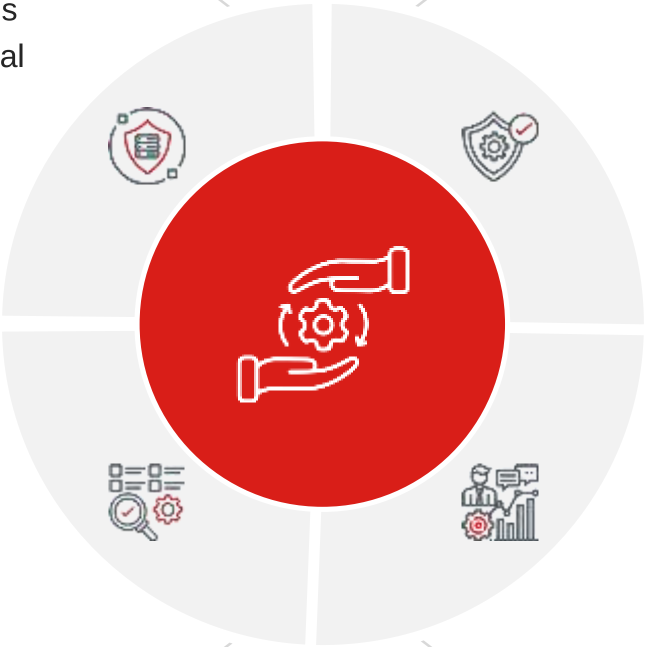
What do we understand by Data Governance?

To Improve Org's trust & reliability in the data and improve usability for Analytics

To allow visibility on data SLA & benchmarks

To ensure compliance with external regulatory requirements

To infuse Data Culture & Data Literacy



Data Quality

- Coverage: RDBMS / NoSQL / Data Lake / File Systems
- Accuracy
- Consistency
- Completeness
- Duplicates & Uniqueness
- Confidence Scores
- Backfills / Historical
- Deduplication

Security and Compliance

- Any regulatory requirements with your data
- GDPR
- PII
- Authentication
- Access Management
- Identity management

Data Discovery

- Data Catalog, Active & Static Meta Data Management
- Data Profiling, Classification, Business Glossary, Data Dictionaries
- ML techniques can help identify facts, trends and relationship among data
- Data Lineage, Reliability, Usability

Data Observability

- Trustworthy and reliable data
- Automated data monitoring, alerting and RCA
- Healthier data pipelines, continuous data monitoring
- Freshness, Distribution, Volume and Schema, Lineage
- Data Drifts, Model Drifts, Proactive avoidance of Decays

Governance Practices

Must have objectives to achieve agreed governance on the data families

Follow FAIR Principles

- Findable
- Accessible
- Interoperable
- Reusable

Data Quality

- Centralized Quality framework
- Config driven self serve Data quality application
- Global and local rules segregation
- Daily quality score update sent to business

Federated governance

- Clear ownership hierarchy with clarity on roles and responsibility
- Distributed ownership structure to achieve MECE

Self-serve data Reporting platform

- One stop shop to view reports.
- Self serve capability to users.

Other Critical Features

- Centralized monitoring of all ETLs
- Implementation of Data Catalogue
- Implementation of Data Lineage

Design principles – FAIR framework

	Findable	Accessible	Interoperable	Reusable
Significance	Metadata and data should be searchable and should be easily located	Metadata and data should be accessible to all relevant users.	Data should be formatted in a way that it can be stored, accessed, processed by multiple applications, and integrated with other data. Additionally, metadata should include qualified references to other metadata.	Metadata should include rich business and technical context. It should be well described so that it can be replicated.
Principles	F1. (Meta)data are assigned a globally unique and persistent identifier F2. Data are described with rich metadata (defined by R1 below) F3. Metadata clearly and explicitly include the identifier of the data they describe F4. (Meta)data are registered or indexed in a searchable resource	A1. (Meta)data are retrievable by their identifier using a standardised communications protocol A1.1 The protocol is open, free, and universally implementable A1.2 The protocol allows for an authentication and authorisation procedure, where necessary A2. Metadata are accessible, even when the data are no longer available	I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. I2. (Meta)data use vocabularies that follow FAIR principles I3. (Meta)data include qualified references to other (meta)data	R1. (Meta)data are richly described with a plurality of accurate and relevant attributes R1.1. (Meta)data are released with a clear and accessible data usage license R1.2. (Meta)data are associated with detailed provenance R1.3. (Meta)data meet domain-relevant community standards
Tools	**TBD**	**TBD**	**TBD**	**TBD**

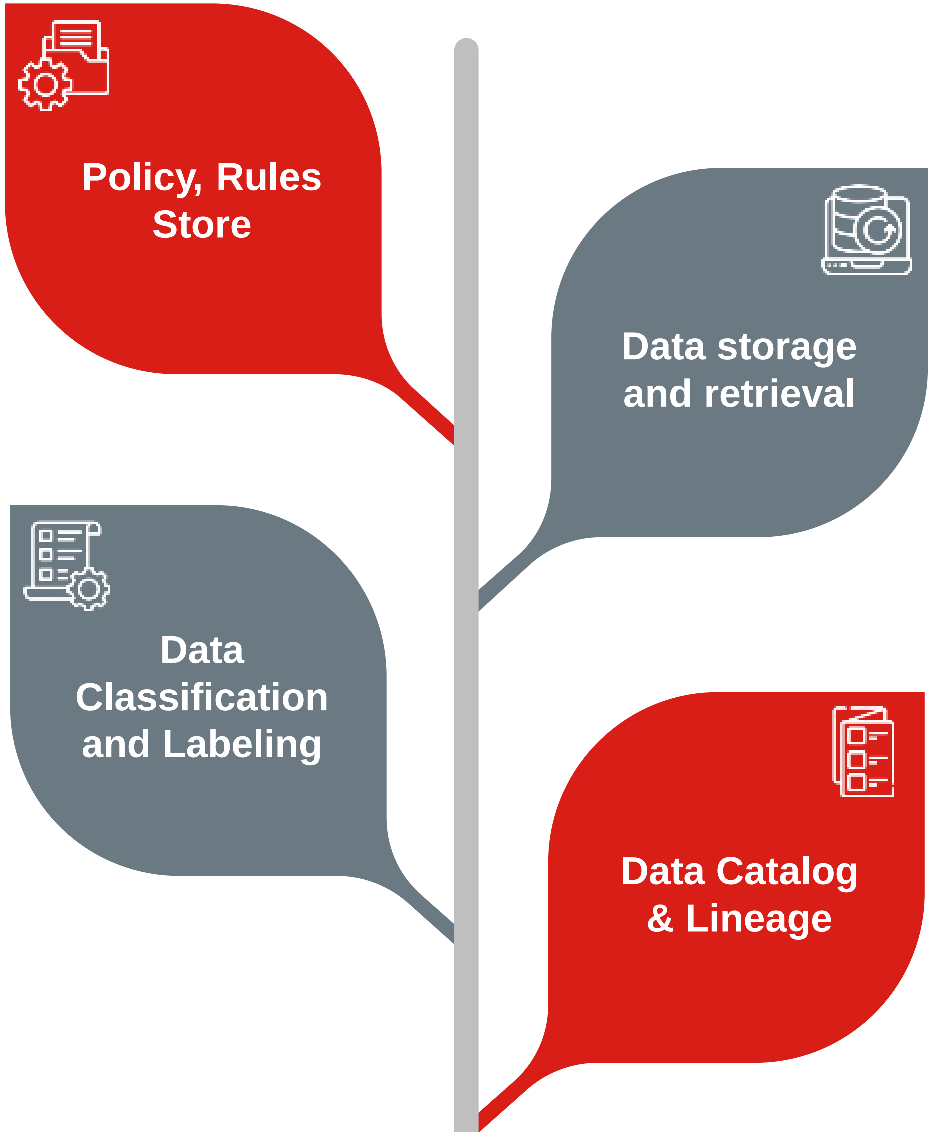
Sigmoid’s Capabilities on Data Governance

What & How

	Use-cases	Lineage	Control Implementation	Monitoring
What we do?	• Identify Data Families • Define Data Family Hierarchy • Data Profiling • Establish Governance Rules • Source Identification • Mapping Creation • Version Controls • Documentation and Validation	• Establish Data Element Priorities • Define Lineage Strategy • Tool Selection • Data Element Documentation	• Identify Control Requirements • Configuration Design • Alert Mechanism Implementation • Testing and Validation	• Policy Development • Roles and Responsibilities • Compliance Monitoring • Training and Awareness
How we do?	• Identify and catalog all data sources feeding into critical data elements & specifying their attributes. • Develop detailed mappings illustrating the transformation logic from source to target data elements. • Implement version control mechanisms to track changes in mappings and ensure accuracy and consistency. • Conduct collaboration sessions with data owners and stewards.	• Identify critical data elements in reports based on business impact and regulatory requirements. • Create a system flow and a logical flow for data lifecycle • Choose appropriate tools (e.g Collibra) for lineage management based on scalability and features. • Document critical data elements' journey from source to reporting, using chosen tools for consistency.	• Assess regulatory, compliance, and internal policy requirements to determine necessary control parameters. • Define control parameters in a way that enables easy configuration at different implementation stages. • Integrate alert systems into the control framework for real-time notification of data quality issues. • Conduct rigorous testing of controls across stages to ensure functionality and adaptability.	• Develop robust control policies aligned with regulatory standards and organizational goals. • Implement regular audits and checks to ensure adherence to the defined policies. • Conduct training sessions to educate stakeholders about data governance policies and procedures.

Data Management

The proposed data management framework recommends centralization policies for data retention and deletion, specific conventions for data classification and labeling, secure data storage and retrieval, and automated data lineage and cataloging (scenario explained wrt Azure tools and services)



Policy, Rules Store

- Azure SQL can act as a centralized store for data management policies.
- Centralization of policies like data retention and deletion.
- Azure SQL will offer high availability, scalability, security, and performance
- Metadata and policies can be stored in Azure SQL to avoid scattered rules local systems.
- Policies changes can be tracked using Azure SQL actions for alerting and notification.

Data storage and retrieval

- Data in production for 5 years before archiving to cold storage
- Access restricted based on user roles for files and reports
- Provisioning file level locks in RDBMS solutions like Azure SQL for data integrity
- Historical data migration may require planning and sourcing from various systems
- Inline reporting, linking to reports, and downloadable custom days with a date range can be considered

Data Classification and Labeling

- Classify and Label data based on sensitivity and importance
- Specific convention for file classification and labeling based on the domain/data family it belongs to

Data Catalog & Lineage

- Automating end-to-end data lineage is ideal for detailed impact analysis.
- IT can troubleshoot and data owners can make relevant changes in the dataset with automated data lineage.
- Cataloging helps ramp up data knowledge for business and data analysts.
- Setting up the data catalog increases data trust, transparency, and compliance.

Data Security

Data Security and Encryption

- Data security and encryption are crucial for protecting the data from unauthorized access and modification
- Secure protocols and encryption algorithms are used to transfer data from sources to cloud storage
- Data at rest is encrypted with keys stored in a separate location from the data
- Regular backups and audits are performed to monitor data quality and security

Data sharing and collaboration

- Data sharing and collaboration are essential for the success a project
- Reports to be shared within the company and outside in accordance with the company's data security policy
- The guidelines aim to ensure responsible and ethical use of data for the benefit of the customer and partners

Data Protection and Remediation

- Use advanced encryption, firewalls, and antivirus software for data security
- Monitor for suspicious activity and alert the Information Security team in case of issues
- Immediately report security incidents to the Information Security team for remediation and prevention
- Maintain usage audit trails for accountability, transparency, and detecting unauthorized access
- Comply with legal and regulatory requirements for data migration, transfer, and archiving

Data Versioning & Quality

Governance Guidelines emphasize the importance of ensuring version control, compliance, security, and flexibility for future regulatory requirements.

01 Version Control and Change Management

- For the changes in the code, data and models managed by the team
- Azure blob storage bucket versioning for data versions.
- Github for collaborative coding and code review
- Azure Databricks versioning for notebook and dataframe changes
- Azure ML versioning for model and dataset changes

02 Regulatory and Compliance Requirements

- The system must comply with regulatory and compliance requirements for data privacy, security, quality, and governance using the best practices applicable as per respective market and location.
- There is currently no PII or direct consumer information in scope, but this might change in the future
- The system must be designed with flexibility and scalability in mind to accommodate changes or updates in the requirements

Data Privacy & Controls

Governance Guidelines emphasize the importance of ensuring the confidentiality, integrity, and availability of data through robust security measures and access controls.

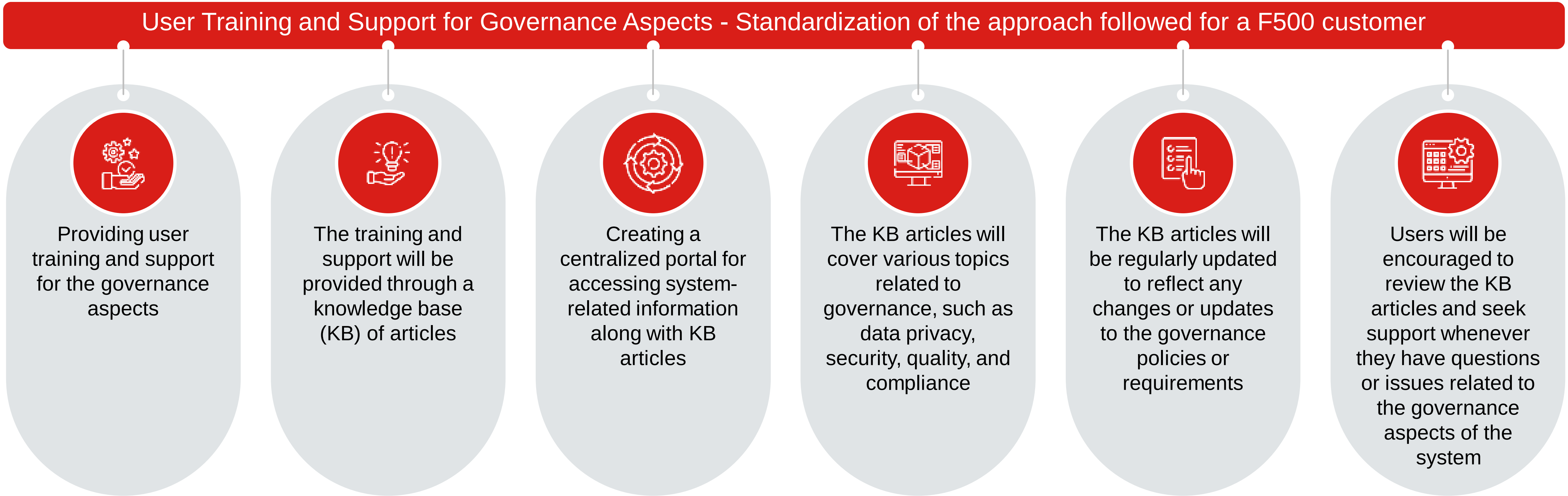
Access Control and Authorization

- Systems to use Active Directory (AD) for access control and authorization
- AD will serve as a corporate directory service and identity provider for users and groups
- Users will be authenticated using AD and granted access to the data store based on their roles and permissions
- External users will use temporary and limited contractor credentials to access the data store
- The system will review and improve access control measures in the future to ensure that only authorized users can access the data store

Data privacy and compliance

- Data privacy policy must be followed to protect personal or sensitive data from unauthorized use or disclosure
- Compliance must be in line with applicable laws, regulations, standards, or policies
- Systems should comply with relevant privacy regulations and maintain validation rules to ensure data quality and accuracy
- Processes for checking data for errors or inconsistencies must be set up and run regularly, and invalid data must be identified and remediated
- Data shared with partners and external entities must be protected, such as by restricting cross-geographic access and utilizing encryption
- Additional authentication measures or granting different levels of data visibility to remote and corporate employees

User Training



User Training and Support for Governance Aspects - Standardization of the approach followed for a F500 customer

Providing user training and support for the governance aspects

The training and support will be provided through a knowledge base (KB) of articles

Creating a centralized portal for accessing system-related information along with KB articles

The KB articles will cover various topics related to governance, such as data privacy, security, quality, and compliance

The KB articles will be regularly updated to reflect any changes or updates to the governance policies or requirements

Users will be encouraged to review the KB articles and seek support whenever they have questions or issues related to the governance aspects of the system

Establish a data-driven culture for operational efficiency

01 Data Democratization

02 Data consistency

03 Develop & revisit data governance policies

04 Implement data security measures

05 Invest in data quality tools and processes

06 Adopt data management tools

07 Foster a culture of continuous learning

08 Establish data ownership and accountability

Success Stories

We have helped implement Enterprise wide DQ and Cataloguing solutions for global clients

Problem Statement	Solution Implemented	Results
• Client's established data pipelines were ingesting and processing more than 1.5 TB of data daily, across business functions in North America and Europe. • The data engineering team dealt with inaccurate or unreliable data which resulted in flawed analyses, jeopardizing strategic decision-making.	A centralized portal with predefined rules to drive data quality across Org with high degree of reusability • Sigmoid deployed a data quality management solution on top of the existing architecture • Seamless integration with diverse data sources, including flat files, CSV files, SQL and No-SQL databases, and other data warehouses. • It was also integrated with automated CI/CD pipelines which improved the overall turnaround time for any code push and enhanced the efficiency of delivery cycles.	• 1.5 TB of data scanned and processed per day • 99% improvement in speed of data quality checks • Automated data diagnostic report generation
A global fast-food chain had consistent issues with quality of data across regions when it comes to sales reporting, supply-chain planning and management reporting (Balance Score Card initiative) Lack of trust in the current data, coupled with acquisition and Integration of of another restaurant chain resulted in need for a global framework	5 Pillars of data check implemented across the functions 1. Data Source Check (Back Office /Vendor Data/ SOURCE DB/S3/SFTP/API) 2. DNA DataLake Check (Data Health) 3. DNA ETL Pipeline Check 4. DNA Data Warehouse Check • Data completeness check • Data quality check 5. DNA Reporting Check (DNA dashboards)	• Standardized framework being rolled out across functions • 82% improvement in data accuracy and and 90% increase in timeliness of data • Automated alerting and ticketing

Engagement Models

Sigmoid’s Engagement Models

Project Based

- Starts with consulting/scoping (2-3 weeks)
- Delivery Program Management
- Interim review
- Success criteria met and IP handover
- Option to continue with product support
- Fixed bid contract
- 3-5 months duration given complexity of problem

Benefits

- Cost effective
- KPI/SLA/Outcome driven
- Suitable for Fixed scope of work
- Less overheads

Staff Augmentation

- Understanding of skill requirements
- Profile match and rate card
- Onboarding and monthly billing
- Focused training based on client tech stack
- Project Management support
- 10% backup resources unbilled and trained

Benefits

- Scalability
- Flexibility in resourcing
- Ability to change/redefine scope

Hybrid-Flexi Model/Data Labs/CoE

- Mix of project and staff augmentation engagements
- Requirement gathering
- Requirement classification - as project or staff augmentation
- Joint delivery plan
- Secure resources internally from Sigmoid and bill monthly
- Dedicated PM, Engineering Managers
- Dedicated Management Consultant(s)
- Dedicated Team Leads and Product Owners

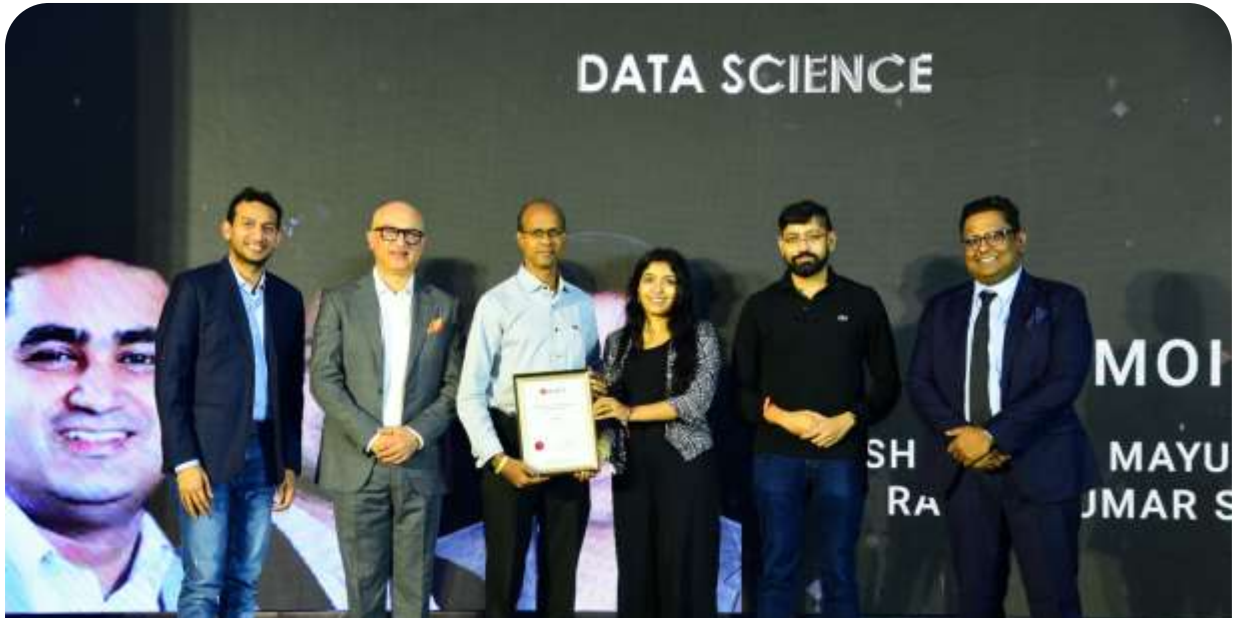
Benefits

- Cost effectiveness by focus on output
- Ability to change/redefine scope/Change requests
- Risk/Reward linked to KPI/SLA

Thank you

Email: surabhi.s@sigmoidanalytics.com

Website: www.sigmoid.com

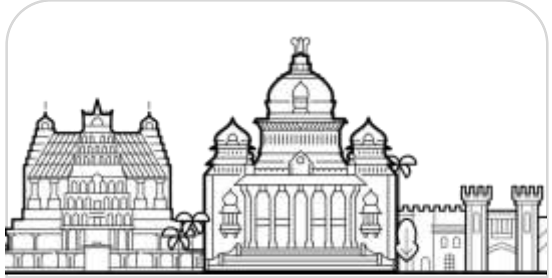


'India Future Unicorn Award'
in Data Science category by Hurun India

Global presence:

USA (NY, SF, Dallas, Chicago)

EU (Amsterdam, London)



India (Bengaluru)

LATAM (Lima)