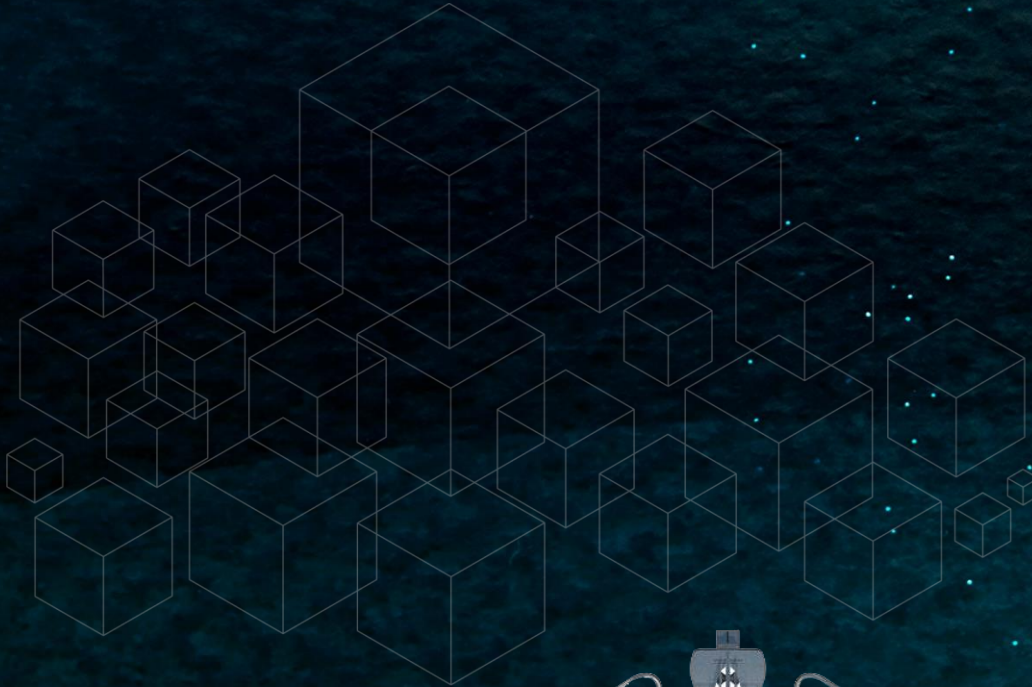


DATA LAKE NAVIGATION

Unraveling Insights in the Sea of Information

A SCALABLE WHITEPAPER

CATEGORY : DATA



SCALABLE
AI

TABLE OF CONTENT

INTRODUCTION.....	3
THE NEED OF DATA LAKE.....	5
KEY BENEFITS OF DATA LAKE.....	6
DIFFERENCES BETWEEN DATA LAKE AND DATA WAREHOUSE?.....	8
DATA WAREHOUSE GAPS FILLED BY DATA LAKE.....	9
DATA SOURCES FOR DATALAKE.....	10
DATA LAKE ARCHITECTURE.....	11
DATA LAKE REFERENCE ARCHITECTURE.....	13
DATA GOVERNANCE.....	14
FUTURE OF DATA LAKE.....	16
CONCLUSION.....	17
ABOUT SCALABLE AI.....	18

INTRODUCTION

Data can be traced from various consumer sources. Managing data is one of the most serious challenges faced by organizations today. Organizations are adopting the data lake models because lakes provide raw data that users can use for data experimentation and advanced analytics.

A data lake could be a merging point of new and historic data, thereby drawing correlations across all data using advanced analytics. A data lake can support the self-service data practices. This can tap undiscovered business value from various new as well as existing data sources. Furthermore, a data lake can aid data warehousing, analytics, data integration by modernizing. However, lakes also face hindrances like immature governance, user skills and security.

One among four organizations already have at least one data lake in production. Another quarter will embrace production in a year. At this rate analyst not only expect this trend to last long but also forecast it to speed up incorporation of innovative data generating technologies in practice. 79% of users having a lake state that most of the data is raw with some portion for structured data, and those portions will grow as they comprehend the lake better.

Managing data is one of the most serious challenges faced by organizations today. The storage systems need to be managed individually, thus, making

infrastructure and processes more complex to operate and expensive to maintain. In addition to storage challenges, organizations also face many complex issues such as limited scalability, storage inadequacies, storage migrations, high operational costs, rising management complication and storage tiring concerns.

There are two major types of data lakes based on data platform. Hadoop-based data lakes and relational data lakes. Hadoop is more usual than relational databases. However, data lake spans both. The platforms may be on premises, on clouds, or both. Thus, some data lakes are multiplatform as well as hybrid.

Though adopting and working on traditional technologies like data mining and data warehousing is important, it is equally important to adopt modern capabilities that not only makes it more evolved but efficient as well. As organizations need to solve challenges at a faster pace, the need has shifted to adopt hybrid methods to explore, discuss and present the data management scenarios. In the present day industries, ideas like data lake to ease data sharing have erupted, but with traditional methods like data warehousing the scope for growth is limited.

A data lake receives data from multiple sources in an enterprise to store and analyze the raw data in its native format. In an industry, data lake can handle data ranging from structured data such as demographic data or semi-structured data such as pdfs, notes, files to completely unstructured data such as videos and images. Using data lake, organizations can dive into possibilities yet to be explored by enabling data management technology to avoid the functional shortcomings. With the advancements in data science, artificial intelligence and machine learning, a data lake could assist with various efficient working models for this industry, industry related personnel as well as specialized capabilities like predictive analysis for future enhancement.

Although data lake is new face and seems to be in a primitive state, many industry giants like Amazon, Google etc. have worked on it. They have processed data in a faster and reliable manner creating a

balanced value chain. For its deployment, administration, and maintenance, lot of efforts has to be instilled. As it is a pool of data from various organization, it has to be governed, secured and be scalable at the same time to avoid it being a dump of unrelated data silos.

This white paper will present the opportunities laid down by data lake and advanced analytics, as well as, the challenges in integrating, mining and analyzing the data collected from these sources. It goes over the important characteristics of the data lake architecture and Data and Analytics as a Service (DAaaS) model. It also delves into the features of a successful data lake and its optimal designing. It goes over data, applications, and analytics that are strung together to speed-up the insight brewing process for industry's improvements with the help of a powerful architecture for mining and analyzing unstructured data – data lake.



THE NEED OF DATA LAKE

A data lake is a centralized data repository that can store a multitude of data ranging from structured or semi-structured data to completely unstructured data. Data lake provides a scalable storage to handle a growing amount of data and provides agility to deliver insights faster. A data lake can store securely any type of data regardless of volume or format with an unlimited capability to scale and provides a faster way to analyze datasets than traditional methods.

A data lake provides fluid data management fulfilling the requirements of an industry as they try to rapidly analyze huge volumes of data from a wide range of formats and extensive sources in real-time.

A data lake has flat architecture to store data and schema-on-read access across huge amounts of

information that can be accessed rapidly. The lake resides in a Hadoop system mostly in the original structure with no content integration or modification of the base data. This helps skilled data scientists to draw insights on data patterns, disease trends, data abuse, insurance fraud risk, cost, and improved outcomes and engagement and many more.

A data lake gives structure to an entity by pulling out data from all possible sources into a legitimate and meaningful assimilation. Adopting data lake, means developing a unified data model, explicitly working around the existing system without impacting the business applications, alongside solving specific business problems.

HOWEVER, WITH EVERY OPPORTUNITY COMES A CHALLENGE. THE CONCEPT OF "DATA LAKE" IS CHALLENGING, THE ATTRIBUTING REASONS BEING

Entities have several linkages across the enterprise infrastructure and functionality. This leads to non-existence of a singular independent model for entities.

It contains all data, both structured and unstructured, which enterprise practices might not support or have the techniques to support.

It enables users across different units of enterprise to process, explore and augment data based on the terms of their specific business models.

Technology should be able to let organizations acquire, store, combine, and enrich huge volumes of unstructured and structured data in raw format and have the potential to perform analytics on these huge data in an iterative way.

Data lake may not be a complete shift but rather an additional method to aid the existing methods like big data, data warehouse etc. to mine all of the scattered data across a multitude of sources opening new gateway to new insights.

KEY BENEFITS OF DATA LAKE

Having understood the need for the data lake and the business/technology context of its evolution, important benefits in the following list:

SCALABILITY

The Hadoop is a framework that helps in the balanced processing of huge data sets across clusters of systems using simple models. It scales up from single server to thousands, offering local computation and storage at each node.

Hadoop supports huge clusters maintaining a constant price per execution bereft of scaling. To accommodate more one just has to plug in a new cluster. Hadoop runs the code close to storage getting massive data sets processed faster. Hadoop enables data storage from disparate sources like multimedia, binary, XML and so on.

HIGH-VELOCITY DATA

The data lake uses tools like Kafka, Flume, Scribe, and Chukwa to acquire high-velocity data and queue it efficiently. Further they try to integrate with large volumes of historical data.

STRUCTURE

The data lake presents a unique arena where structure like metadata, speech tagging etc. can be applied on varied datasets in the same storage with intrinsic detail. This enables to process the combinatorial data in advanced analytic scope.

STORAGE

The data lake provides iterative and immediate access to the raw data without pre-modelling. This offers flexibility to ask questions and seek enhances analytical insights.

SCHEMA

The data lake is schema-less write and schema-based read in the data storage front. This helps to develop up-to-date patterns from the data to grasp applicable intelligent insights without being dependent on the data.

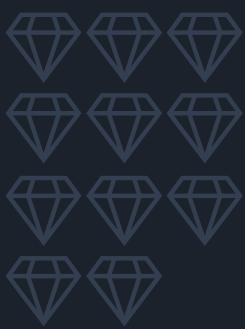
SQL

Pre-existing PL-SQL scripts could be reused once the data is stored in the SQL storage of the data lake. The tools like HAWQ and IMPALA give flexibility to process huge parallel SQL queries while working in parallel with algorithm libraries like MADLib and SAS applications. Performing the SQL takes less time and also consumes less resources inside the data lake than performing outside.

ADVANCED ALGORITHMS

The data lake is proficient at using the large amount of understandable data along with advanced algorithms to acknowledge article of interest to power up decision making algorithms.

Data Lake Embraces NEW DATA SOURCES



DATA LAKE

60%
Global Data
Growth
Per Year

2.7
Zetabytes
Data
In Digital
Universe

Data Lake Embraces
NEW DATA SOURCES

80% Industry Information
is Unstructured

42% Businesses Admit Unstructured
Data Hard to Interpret

Data Lake Helps Business Organizations.
Capture, Manage and Analyze all Their Data.

DATA LAKE



DATA LAKE

Data Lake Reduces
THE COST OF DATA

Up to 20% IT Spending on
Data Storage



Required to Upgrade
Legacy Systems

Data Lake Eliminates
TIME CONSTRAINTS

61% Industries Want
Faster Access to Data



Reports can Take
Days or Weeks

DIFFERENCES BETWEEN DATA LAKE AND DATA WAREHOUSE?

James Dixon's idea of a new architecture known as 'Data Lake' developed in 2010, gained quite a bit of momentum and captivated numerous data driven industries. It is easily accessible and can store anything and everything indifferent of its type, structure etc. Data lake and data warehouse have different but valuable characteristics to offer to an enterprise together. The major differences between a data lake and a traditional data warehouse are:

Clearly data lake and data warehouse are complementary to each other in an/the enterprise. Data lake should not be seen as a replacement for a data warehouse as they are unique in their own way, having distinct roles in the industry.

DATA LAKE	DATA WAREHOUSE
A data lake can store structured, semi- structured as well as unstructured data	Data warehouse only accommodates structured data that is given in a particular model
Data lake contains relevant data, easy to access and provides operational back-up to the enterprise	Data warehouse stores data for longer period of time which can be accessed on demand
In a data lake, data doesn't need to be modeled and only raw data needs to be loaded and used	Data in data warehouse needs to be modeled before processing data and loading it
Data lake has enough processing power to analyze and process data being accessed	Data warehouse only processes structured data into a reporting model for reporting and analytics.
A data lake is easy to reconfigure	Data warehouse structure is difficult to reconfigure because the data is highly structured
Data lake costs less to store data, doesn't need licensing and is built on Hadoop, an open source framework	With data warehouse, optimization is time consuming and is costly. This works well with pre-existing modeled data but falls flat for new data.
In data lake, data availability can be easily spotted and integrated for a requirement. Back tracking of data and data management are available and easy to implement.	With data warehouse, data availability is tough to spot and integrate for a particular requirement. Back tracking of data and data management are unavailable and cumbersome to implement. Manual creation of root data is error prone and time consuming.

DATA WAREHOUSE GAPS FILLED BY DATA LAKE

Data lake supports multiple reporting tools and has a self-sufficient capacity. It helps elevate the performance as it can traverse huge new datasets without heavy modeling. –Flexibility. It supports advanced analytics like predictive analytics and text analytics. This further allows users to process the data to track history to maintain data compliance. – Quality. Data lakes allow users to search and experiment on structured, semi-structured, unstructured, internal, and external data from variable sources from one secure viewpoint. – Findability and Timeliness.

There are many challenges to overcome in the data warehouse. The solution that suffices all of the gaps in the data lake. This helps to secure the data and work on the data, run analytics, visualize, and report on it. The characteristics of a stable data lake are as follows:

USE OF MULTIPLE TOOLS AND PRODUCTS

Extracting maximum value out of the data lake requires customized management and integration that are currently unavailable from any single open-source platform or commercial product vendor. The cross-engine integration necessary for a successful data lake requires multiple technology stacks that natively support structured, semi-structured, and unstructured data types.

DOMAIN SPECIFICATION

The data lake must be tailored to the specific industry. A data lake customized for biomedical research would be significantly different from one tailored to financial services. The data lake requires a business-aware data-locating capability that enables

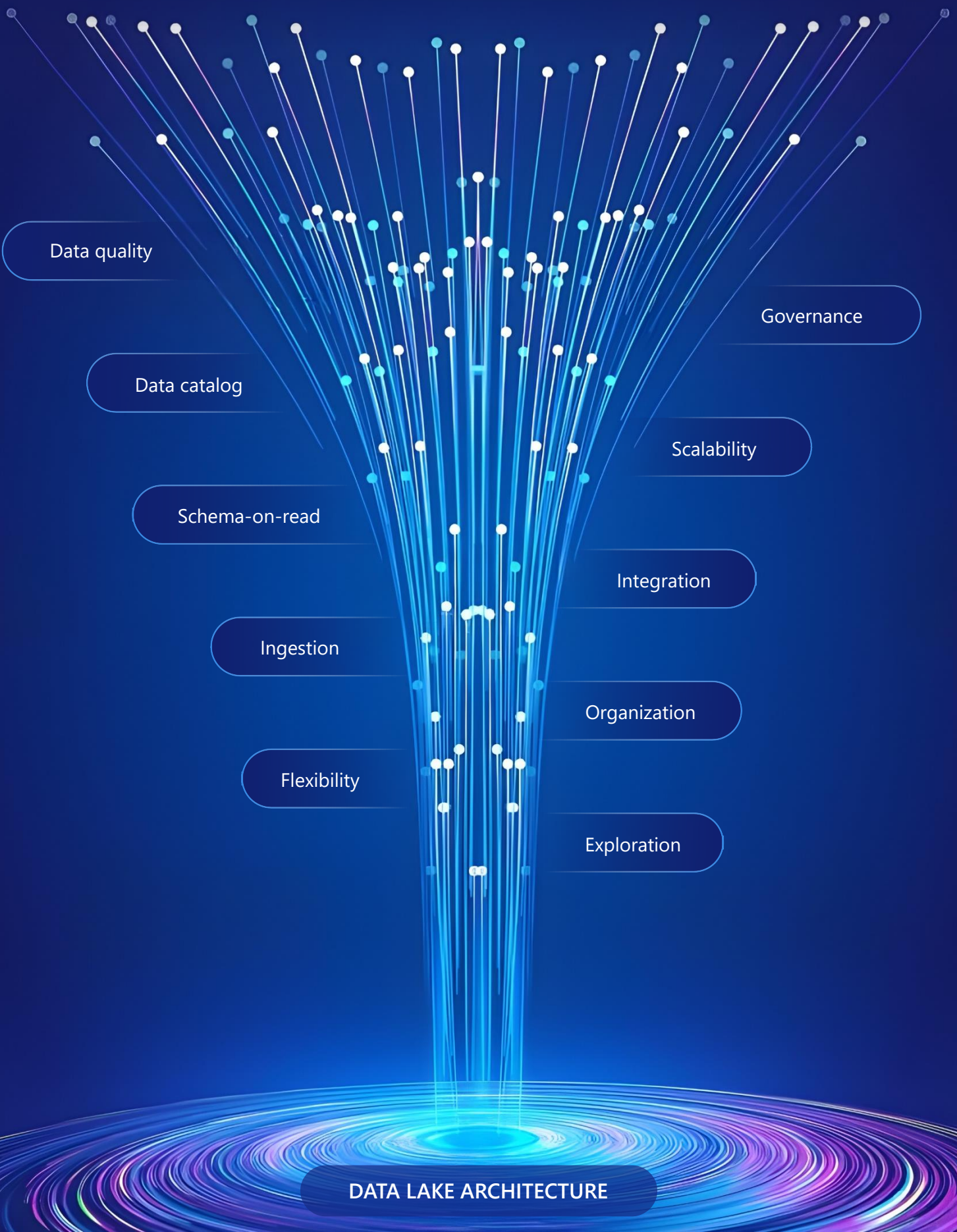
business users to find, explore, understand, and trust the data. This search capability needs to provide an intuitive means for navigation, including key word, faceted, and graphical search. Under the covers, such a capability requires sophisticated business ontologies, within which business terminology can be mapped to the physical data. The tools used should enable independence from IT so that business users can obtain the data they need when they need it and can analyze it as necessary, without IT intervention.

AUTOMATED METADATA MANAGEMENT

The Data Lake concept relies on capturing a robust set of attributes for every piece of content within the lake. Attributes like data lineage, data quality, and usage history are vital to usability. Maintaining this metadata requires a highly-automated metadata extraction, capture, and tracking facility. Without a high degree of automated and mandatory metadata management, a data lake will rapidly become a data swamp.

CONFIGURABLE INGESTION WORKFLOWS

In a thriving data lake, new sources of external information will be continually discovered by business users. These new sources need to be rapidly on-boarded to avoid frustration and to realize immediate opportunities. A configuration-driven, ingestion workflow mechanism can provide a high level of reuse, enabling easy, secure, and trackable content ingestion from new sources.



Data quality

Governance

Data catalog

Scalability

Schema-on-read

Integration

Ingestion

Organization

Flexibility

Exploration




DATA LAKE ARCHITECTURE

DATA LAKE ARCHITECTURE

Data lake architecture should be flexible and organization specific. It relies around a comprehensive understanding of the technical requirements with sound business skills to customize and integrate the architecture. Industries would prefer to build the data lake customized to their need in terms of the business, processes and systems.

An evolved way to build a data lake would be to build an enterprise model taking few factors into consideration like, organization's information systems and the data ownership. It might take effort but provides flexibility, control, data definition clarity and partition of entities in an organization. Data lake's self-dependent mechanisms to create process cycle to serve enterprise data, help them in consuming applications.




The Data lake as composed of three layers and tiers. Layers are the common functionality that cut across all the tiers. These layers are:

-  DATA GOVERNANCE AND SECURITY LAYER
-  METADATA LAYER
-  INFORMATION LIFECYCLE MANAGEMENT LAYER

-  INTAKE TIER
-  MANAGEMENT TIER
-  CONSUMPTION TIER

Tiers are abstractions for a similar functionality grouped together for the ease of understanding. Data flows sequentially through each tier. While the data moves from tier to tier, the layers do their bit of processing on the moving data. The following are the three tiers:

The Data lake as composed of three layers and tiers. Layers are the common functionality that cut across all the tiers. These layers are:

-  CONSISTENCY
-  AVAILABILITY
-  PARTITION TOLERANCE

DAaaS (Data Analytics-as-a-Service) is a protractible platform. It uses a cloud-based delivery model. It provides a wide range of tools to select from for data analytics that can be designed by the user to process large amounts of data effectively. Enterprise data is ingested into the platform. Further the data is processed by analytics applications. This could provide business insight using advanced analytical algorithms and machine learning

As per researchers, experts and data enthusiasts, the "Data Lake" to "a successful Data and Analytics" transformation needs the following:

DAAAS STRATEGY SERVICE DEFINITION

Our Informationists leverage define the catalog of services to be provided by the DAaaS platform, including data onboarding, data cleansing, data transformation, datapedias, analytic tool libraries, and others.

DAAAS ARCHITECTURE

We help our clients achieve a target-state DAaaS architecture, including architecting the environment,

selecting components, defining engineering processes, and designing user interfaces.

DAAAS POC

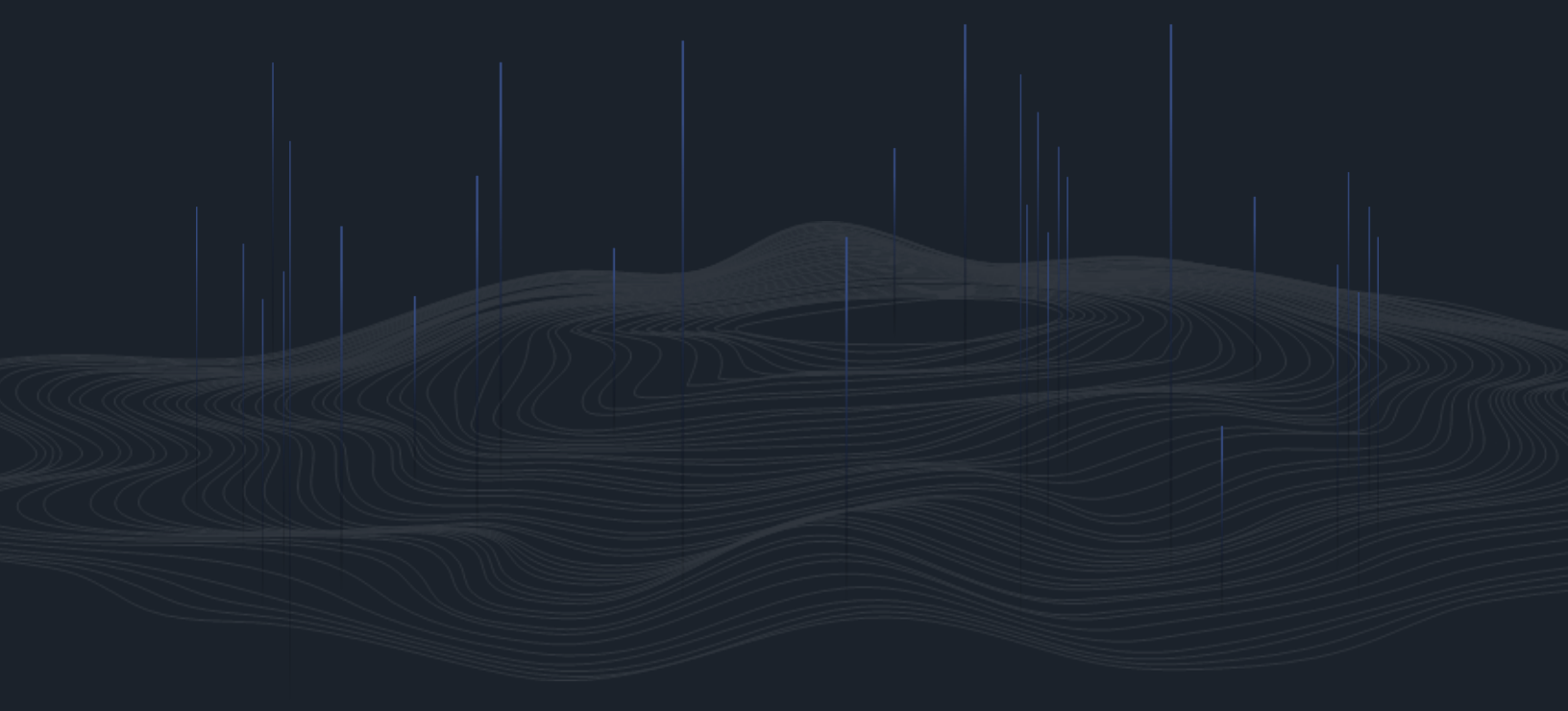
We design and execute Proofs-of-Concept (PoC) to demonstrate the viability of the DAaaS approach. Key capabilities of the DAaaS platform are built/demonstrated using leading-edge bases and other selected tools.

DAAAS OPERATING MODEL DESIGN AND ROLLOUT

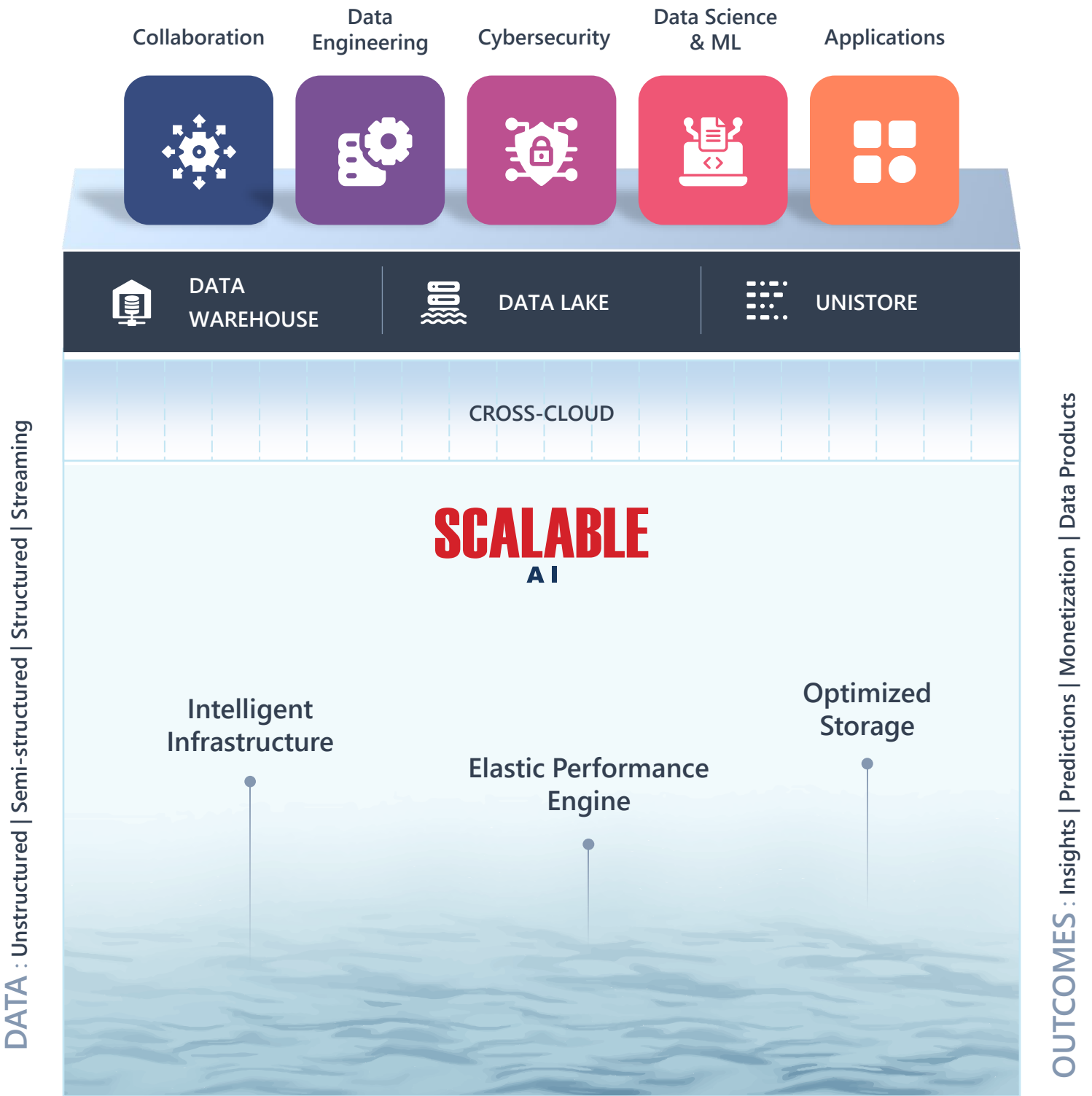
We customize our DAaaS operating models to meet the individual client's processes, organizational structure, rules, and governance. This includes establishing DAaaS chargeback models, consumption tracking, and reporting mechanisms.

DAAAS PLATFORM CAPABILITY BUILD-OUT

We provide the expertise to conduct an iterative build-out of all platform capabilities, including design, development and integration, testing, data loading, metadata and catalog population, and rollout.



DATA LAKE REFERENCE ARCHITECTURE



DATA GOVERNANCE

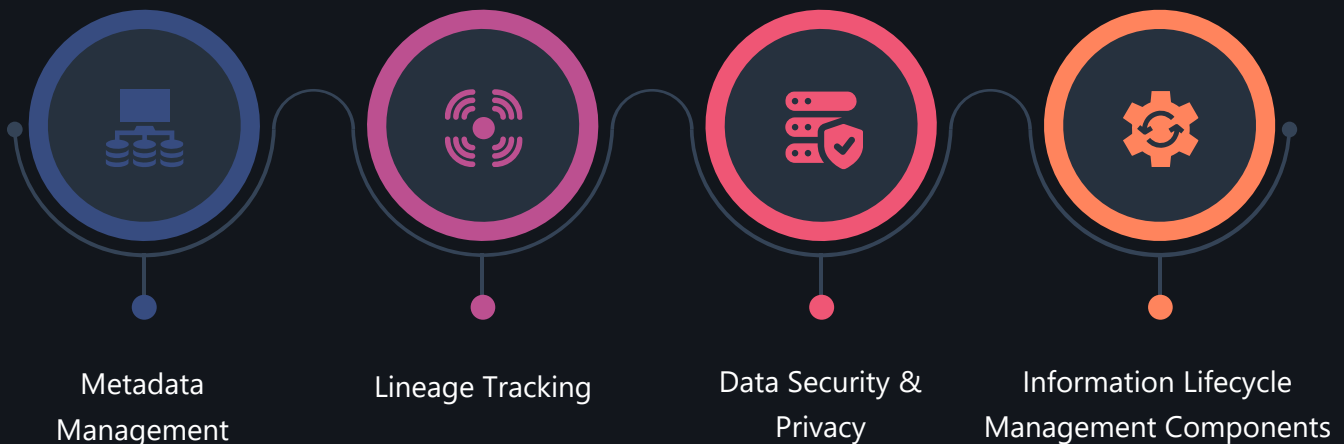
Data Governance is discipline that is enforced on data by an organization as it moves from input to output, making sure that it is not rigged in any way that is risky. To meet the strategic goals, an organization has to convert ingested data into intelligence on a fast pace and accurate basis. It strengthens the decision making process since the data is adhering to certain quality standards. This has a huge effect in enhancing the final value of data, enabling optimal performance planning by data management staff and minimizing rework.

Data Governance deals with processes to lay down the technology architecture to help store and manage mass data. Further, Data Governance deals with the right security policy of the data as it is being acquired and as it flows through the enterprise. While the data is worked upon to derive

a new form, Data governance assures the integrity and accuracy is not meddled with. To maintain and prevent shortage of storage space, data past its usage date is moved to tape storage or is defensively destroyed. This process is owned by Information Lifecycle Management policies, a subset of Data Governance processes.

Organization without a strong Data Governance process end up jeopardizing the caliber of analytics and decisions deduced from it. This exposes the organization to a substantial risk. On the other hand, organizations having a strong built Data Governance processes have arrangements to improve Data Security with intrinsic authentication and authorization in place. In addition, they also have data loss guarding systems.

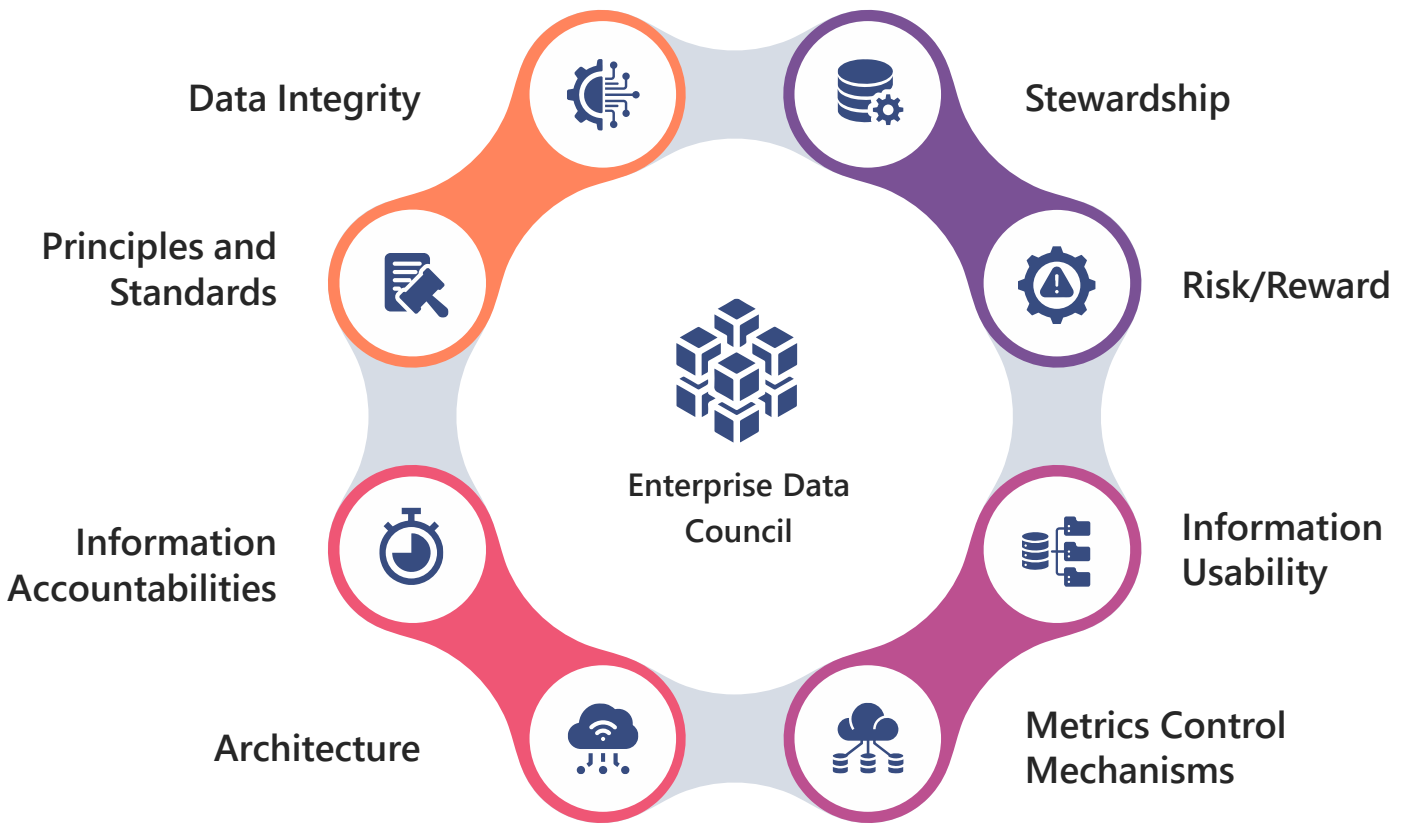
The basic components of Data Governance that cuts across the Data Intake, management, and consumption tiers of the Data Lake are:



Data Steward Project Team	Governance Committee	Enterprise Data Council
<ul style="list-style-type: none"> ● Acts on Requirements ● Build Capabilities ● Does the Work ● Responsible for Adherence 	<ul style="list-style-type: none"> ● Highlights Work ● Drives Change Accountable for results 	<ul style="list-style-type: none"> ● Executive Oversight



← VALUE CREATION →



← ORGANIZATIONAL ALIGNMENT →

FUTURE OF DATA LAKE

The Data Lake deals with the storage, management, and analytical aspects related with the facets of Big Data. Monitoring the Big Data using the Data Lake

adhering to the existing data governance methods to deal with Big Data and also cut a space for expansion of emerging trends.

In the future, it is expected that organizations' data analytics methodologies would evolve to fetch the following abilities:

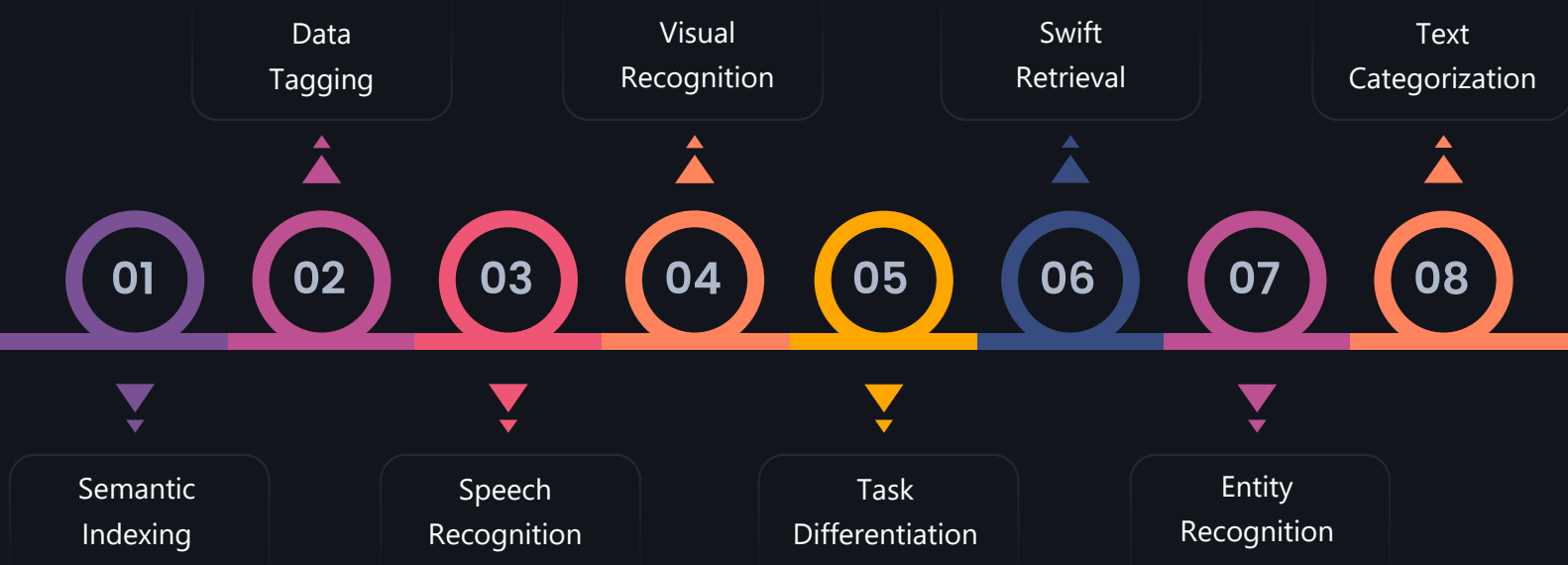
The demand for extremely high-speed insights

The growth of Internet of Things

The adoption of cloud technologies

The evolution of deep learning

DEEP LEARNING IS TYPICALLY USED TO SOLVE SOME OF THE INTRACTABLE PROBLEMS OF BIG DATA AS FOLLOWS:



Data Lake possesses the optimal balance to enable organizations to tap the real benefits of advanced analytics and Deep Learning methods. The Data Lake property to store data in a schema less way is immensely useful for the methods to extract

unstructured data representations. Data Lake also has the ability to whisk complicated high end deep learning algorithms on elevated-dimensional and streaming data.

CONCLUSION

Data lakes with advanced analytics are reshaping the way enterprises work. Future with data lakes looks very promising. System developers are immersed in vigorous R&D for such technology advancement for better analysis and detail oriented search. It could be useful for industries by providing better efficiency and outcomes.

To be at an advantage, industry will have to use the power of data lake driven processes and systems. If fathomed intuitively, it could change the way services is being delivered.

Presently, data lake practices are governed by Hadoop predominantly. Hadoop has become the major tool for assimilating and pulling out insights from combinatorial unstructured data present in Hadoop and enterprise data assets, running algorithms in batch mode using the MapReduce

paradigm. Hadoop, with the existing enterprise data assets such as data in mainframes and data warehouses. Languages such as Pig, Java Map Reduce, SQL variants, RHadoop, Apache Spark, and Python are being increasingly used for data munging, data integration, data cleansing, and running distributed analytics algorithms.

There is more to consider with details including: big data architecture for accessible Data Lake infrastructure, data lake functionality, solving data accessibility and integration at enterprise level, data flows in the data lake, and many more. With these numerous queries, there still is resources to tap and a lot to gain for the enterprise. Using the data lake architecture to derive cost efficient, life-changing insights from the huge mass of data nullifies the concern regarding going further with the ice-berg hidden under the ocean.

About Scalable AI

We deliver actionable insights that organizations can use to identify opportunities, manage risks, achieve operational excellence, and to gain an innovative edge.

www.scalableai.com

About Scalable Systems

Scalable Systems is a Data, Analytics & AI Company focused on vertical-specific innovative solutions. By providing next-generation technology solutions and services, we help organizations to identify risks & opportunities, and achieve sales and operational excellence to gain an innovative edge.

www.scalable-systems.com