Nuvalence

ARCHITECTURE GUIDANCE SERIES

The Anatomy of a Data-Driven Digital Platform

WHITEPAPER

The Anatomy of a Data-Driven Digital Platform

Purpose & Audience

Converting raw data into meaningful insights is challenging enough for most organizations, but generating better decisions from data continuously at scale is the road less traveled. While more and more data are being generated every day, transforming and converting that data into material decisions is a complex and nontrivial journey filled with a multitude of constraints. The variety of data formats, the increasing degrees of data diversity, and the ever surging volume of generated data make it difficult to define repeatable patterns for all industries and domains. The purpose of this paper is to describe the questions your organization needs to ask and answer in order to navigate business and technical resources toward data-driven insights. We then propose a conceptual architecture of shareable components meant to accelerate data processing and analysis across teams, otherwise known as a "data-driven digital platform."

The business or mission justification behind a data-driven digital platform (or moresimply a "data platform") is to generate meaningful insights and decisions within a common domain (e.g. automotive, healthcare, banking, retail). Ideally, by making it easier to explore and understand data, a platform can open up the possibility to new digital outcomes and products, potentially ones that were previously unforeseen prior to embarking on a data platform journey. A platform that makes it easy for tenants (e.g. engineers, analysts, data scientists, et al.) to discover and consume platform resources, the more likely their work will lead to new and greater decisions that benefit the organization. This paper is intended for Business Stakeholders, the Chief Data Officer (CDO), Chief Data Scientists, and anyone else who is looking to create repeatable business or mission successes from acquiring, processing, and analyzing their data.

Understanding the Context

From the Nuvalence Whitepaper, **"The Anatomy of a Digital Platform,"** a digital platform is characterized as a unique taxonomy when compared to the objectives and design of a digital product. Literally speaking, it is "best described as a domain-specific platform...that abstracts common, high-value functions in a technology ecosystem into a shareable set of data, APIs, components, and capabilities."

Data are a natural resource of any digital platform as both people and services generate, store, and consume and consume multiple types of data. With that said, data may be the primary catalyst for commoditizing patterns into a platform architecture for the purpose of accelerating the delivery of data-driven outcomes. Furthermore, a domain-specific data platform is a type of digital platform that helps a variety of consumers and services to acquire data, prepare it for analysis, and eventually derive insights that propel insights for an organization and/or user base.

The **benefits of taking a data platform approach** to organizational priorities and objectives should include:

1. Facilitating functions and services for data sharing and data blending from as many sources as needed in order to unlock potential insights.
2. Extending decisions into production so that the business or the mission recognizes some measurable outcome from the data analytics lifecycle (e.g. better patient care, new market penetration, predicting expected revenue, or mitigating fraud).
3. Scaling out data-driven resources, services, and decision-making assets to multiple constituencies (e.g. software developers, new lines of business, customers, or 3rd party partners).
4. Unlocking unspecified or unexpected use cases and analytic methods for future decision-making (i.e. questions you did not know you wanted to ask until the platform helped expose and reveal new opportunities).

While many other data-driven narratives tend to focus on a particular component of the data analytics lifecycle (e.g. storage, data processing, business intelligence, machine learning), this paper examines how a platform approach can facilitate each component of the lifecycle and accelerate decision-making for multiple tenants and use cases within the domain. The entire purpose of a data-driven platform is getting to more and better business/mission decisions, faster.

As a technology layer at a practical level, a data platform should display the following characteristics:

1. Cleanses, denormalizes, and integrates data from disparate sources into some type of centralized, accessible repository for analytics, reporting, and/or business intelligence (BI).
2. Advances beyond traditional analytics toward automated and statistical learning particularly around larger data sets. Machine Learning (ML) and Deep Learning (DL) have become the de facto scientific and engineering standards when applying Artificial Intelligence (AI) to data.
3. Exposes data-driven discoveries and/or services, which accelerate the delivery of a spectrum of analytics, reporting, or training/serving machine learning models through APIs, user interfaces, cloud services, or direct access to raw data.
4. Provides flexibility for uncovering new and potentially unforeseen use cases derived from access to new data or the blending of data sources, unlocking a diverse set of possibilities and outcomes.

In addition to describing what defines a data platform, it can also be helpful to describe what it is not. A data “product” or set of “services” like a BI engine/dashboard(s), data pipeline(s), data warehouse, or even machine learning model can assuredly provide business and mission value independently. However, if their constituency and use case is limited in scope (e.g. sales forecasting, marketing analytics, profit/loss visualization), and does not scale into additional outcomes then they do not compose the many characteristics indicative of a platform.

Characteristics 3 and 4 are where individual data services and analytic techniques evolve into a platform; where data and insights are exposed, scaled, potentially monetized, and new data use cases cascade into a set of sharable components that continue to deliver positive results. The rest of this paper examines the requirements of characteristics 3 and 4 and then illustrates a common architecture that articulates the components and purpose of a data platform.

Requirements

While the "pieces" that can comprise a data platform can be numerous, it is the variety and sometimes overlapping nature of techniques, disciplines, and vernacular around data that lead to single-use case systems, long stretches of time between data discoveries, or just outright confusion.

Choosing the right components for a data platform depends on organizational objectives, the questions that stakeholders are looking to answer using data, and whether the analytical methods or formulas for answering these questions are known (e.g. a Key Performance Indicator or business metric already defined by the organization) or unknown (e.g. a yet-to-be-trained machine learning model).

Generally speaking, all data platforms will require some consistent set of components for data acquisition, processing, storage, analytics, and exposure. Rather than starting with questions about your platform, you should start with questions and requirements around your data and the mechanisms you can use to glean insights from it. Your answers will help you determine the shape of your platform (at least where to begin on your platform delivery journey). Let us conduct a requirements analysis using Characteristics 3 and 4.

Exposes data-driven discoveries and/or services, which accelerate the delivery of a spectrum of analytics, reporting, or training/serving machine learning models through APIs, user interfaces, cloud services, or direct access to raw data.

1. **What are the business/mission objectives you would like to achieve from understanding your data?**

Multiple stakeholders coupled with large data sets from a myriad of sources tend to lead an organization into numerous and sometimes conflicting objectives. Hypothetically speaking, some business owners may want to find ways to increase revenue by penetrating new markets while others want to understand what drives current customer behavior. Some stakeholders may want basic business reporting while others are looking for new ways to predict trends using different mechanisms.

Data scientists and machine learning engineers (i.e. those with a stake in making predictions using large data sets) may want to focus on the prediction accuracy of a model, whereas **software engineers** may want to

reduce the number of bugs in the platform or increase the performance of the entire system. This is all normal, and different objectives can be addressed in separate, concurrent streams of work. However, some cannot, and it is important to continuously map your objectives from your data to ensure alignment amongst all stakeholders. Roadmap planning and backlog prioritization exercises should pay particular attention to technical outcomes that overlap with more than one business objective.

The aforementioned objective examples do not need to be mutually exclusive when it comes to the taxonomy of a data platform. There are many reusable components like data warehouses and machine learning services that if utilized to answer the right questions at the right stage of the data lifecycle can lead to continuous outcomes. The original Digital Platform Whitepaper by Nuvalence discusses principles such as Tenancy, Entitlements, and the concept of an "Anchor App" to be the initial components to fuel follow-on use cases and new digital products. The same can be the case for a data platform where the "Anchor App" can be a new dashboard, trained machine learning model, or data-driven API endpoint that sparks continuous delivery and discovery using the same, shareable platform resources like data pipelines, data warehouses, dashboard templates, and ML pipelines.

The alternative approach is building systems catered to a single-use case. This can lead to less of a focus on collaboration and reusability across teams, more duplication of work, and if the questions change, it can lead to empty platforms.

2. **Is your method or formula for answering your objectives known and/or unknown?**

A "known" method or formula is something already defined by business stakeholders like a Key Performance Indicator (KPI) or business metric that you routinely want to track and/or evolve over time. BI analysts use existing methods to determine trends, generate reports, and provide input to leaders in charge of defining new methods.

An "unknown" method is something you have yet to define, but are looking to develop through techniques like machine learning. In fact, machine learning is ideal for discovering far more precise and sometimes nonobvious methods over larger data sets than traditional analytics. You can think of machine learning as "automation towards advanced method creation" because it can help define more intelligent formulas for your data in a quicker amount of time than manual analysis, especially when utilized in production.

Start with the outcome you are looking for and build your way back. From this point forward, feedback should be in a continuous loop and should begin with asking, **"what results or outcomes do I want from my data?"** Developing a categorization or even a **"tree of questions"** will help lead to more specific investigations like **"Do I want to predict sales for next year based on data from previous years? Do I want to recognize patterns or better classify information like determining if an insurance claim is legitimate or fraudulent?"**

A more specific way to look at the question in this section is to investigate if the platform leans more toward traditional BI analytics and reporting, or does it lend itself to advanced analytics (i.e. AI/ML/DL), or both?

The drivers for all analytics platforms share a common goal: better insights and decisions through the analysis of data. With the advancement of scalable cloud resources and managed services that make data science more democratized, machine learning has made it easier to automate the discovery of insights in larger and larger amounts of data. It by no means is a requirement for a data-driven platform to have an AI component. Nevertheless, the trend is pointing toward more and more data-driven platforms making use of machine learning as data sets increase in volume, and cloud services make machine learning easier with pre-trained models, low-code interfaces, and automated ML pipeline delivery.

Another misconception is that BI and AI are different, and that makes them mutually exclusive, which is often not the case. Traditional analytics eventually lead to advanced analytics because you have more data and you are looking to automate analysis and method discovery. The relationship can be sequential, but it is also best if it follows a continuous feedback loop. Data analysts use structured analysis and traditional reporting to analyze data about their domain, and it may very well be the starting point for extracting meaningful features for machine learning that results in better labels and parameter tuning during model training.

3. **What are your means for acquiring data?**

The entire data lifecycle begins with an understanding of data locations, data formats, the current state of data quality, the variety of sources that generate your data, and arguably most importantly, the current value of your data to your business. While the objective of a data platform is to continuously uncover more value over time, today's metrics serve as benchmarks for measuring the results of future performance.

Acquiring more data may seem like an initial step, but it is critical to ensure you are making the best use of the data you currently have. Otherwise, you open yourself up to continued poor behaviors and rising storage and processing costs with no existing recipe for creating results. Not to mention, if you are able to provide insights from existing data, this can inform the data you want to collect and it gives validation to the platform approach to the data lifecycle.

Data acquisition can come in many forms that can include transactions, API calls, messages, event stream processing, or batch/micro-batch extraction to name some of the more popular types. Regardless of how you acquire it, gathering, aggregating, and preparing data in a way that is consumable constitutes a significant amount of engineering resources in any data lifecycle, and it is a natural location where most platform components are focused. The reasons for this are numerous: You may not even have all the data you need to inform ideal decisions. Or, as an example, you may have purchased data from a 3rd party, but it is in a repsitory or format that is not easily discoverable, processable, and analyzable by business analysts, datascientists, software developers, or leadership.

A growing challenge most organizations face that makes it difficult to create an acquisition strategy is the amount of data required to train machine learning models. Without enough data or properly processed data, models can all too easily become littered with errors, erode confidence, or result in bias that leads to harmful or offensive outcomes.

The important takeaway from this question is to analyze and enumerate the approach to data acquisition and integration, and measure the value of your sources (e.g. what decisions came as a result of a single data pipeline). Architectural workflows and component descriptions are included in subsequent sections for common methods of data collection.

4. **What is the quality of your data?**

Evaluating data quality is another initial step toward understanding the value of your data regardless of whether it will be useful now or at some point in the future. Your level of data quality is also an indicator for which resources you need to put toward data engineering and data architecture, which could realistically be responsible for the bulk of the work in your data platform. No other component matters without reliable, quality data.

There are many categories of data quality that include: data duplication, missing data, data timeliness, and accurate data to name a few. By measuring and continuously monitoring your data quality you will ensure you are devoting energy to data platform foundations.

5. **What is your level of data diversity?**

In this context, highly diverse data means the many directions, variations, and formats of your data. You will sometimes see comparisons of structured vs. unstructured vs. semi-structured data as ways of describing diversity, and while this is important to examine, there are too many data formats and varieties to consider across all types of data architectures. The more data formats, and overall data volume, the more methods you will need to make use of that data, and the more likely you will need to utilize machine learning and deep learning to help you make sense of it all.

Rather than enumerate all data types, it is more important to illustrate that data sets will need to become more structured and less diverse to be consumable by either a human or a machine. Extract, Transform, Load (ETL) or Extract, Load, Transform (ELT) are common techniques in many environments. For machine learning, data scientists almost always take further steps to process data for model training. Feature engineering and data labeling are common and popular techniques that are quickly becoming a standard for identifying measurable inputs, or features, in your data for model training, validation, and evaluation.

6. **Can you enumerate all of the consumers of your platform and your data (i.e. whoever and whatever needs data)?**

This is a helpful exercise because the more value your data have, the greater the chances that consumers will

make use of it, and the likelihood the variety of consumers will increase over time. If the platform effectively acts as a vehicle for abstracting access to your data, it can funnel data-driven value to consumers in a variety of ways. When discussing consumer personas, several blogs and books address these differences in detail, and many sources (especially job postings) do not always agree on mutually-defined roles and responsibilities.

Sometimes, the consumer is not a person, but an application reading data or calling a model endpoint to make a prediction. Regardless of how the application is using data, it is still likely using it because a person programmed it to do so (at least this is still typically the case in 2022!). For the purposes of this paper, we are focusing on enumerating the data consumer personas.

The table below describes the variety of data consumers in your organization and includes their predominant role in utilizing methods or formulas for meeting the goals of the business. We chose to differentiate whether their primary responsibility is to make use of already predefined analytic methods, or if it is to create new and previously unknown ways of looking at the data.

Title	Role in Meeting Business Objectives	Description
Business Stakeholder	Creates Method	Ultimate decision-maker on business/mission priority to include what types of questions the organization should be asking from their data.
Business Intelligence Analyst	Uses Method	Analyzes domain-specific, post-processed, and structured data to answer questions through existing methods like fact-based reporting and data models.
Data Scientist	Creates Method	Examiner of multiple approaches such as feature extraction and training machine learning models for addressing questions that have never been asked or answered before.
Machine Learning Engineer	Creates Method	Responsible for delivering trained models into a production environment through model testing and assurance, DevOps, Site Reliability Engineering (SRE), and API structure and design.
Data Engineer	Uses Method	Collector and aggregator of data into a more automated and structured format (often via data pipelines) for other consumers like BI analysts and data scientists.
Data Platform Architect	Uses Method	Defines and designs the entire data platform framework that leads to more structured and usable formats of data for multiple consumer types and constituencies.
Data Analyst	Uses Method	Analyst of production machine learning models providing predictions or classifications using large data sets. Monitors...

		model accuracy over time and provides feedback and input for updating model versions.
DevOps	Uses Method	Responsible for providing continuous and often automated delivery of data pipelines, data-driven applications, dashboards, and models into production with a high-quality assurance of the entire system.`
Software Engineer	Uses Method	Builds user experiences and application integrations that access and expose data resources.

NOTE that it is common in many organizations for more than one role outlined above to be owned by the same person. This may make sense given the overlap of data collection, preparation, model development, and model serving. This is one of the reasons the term "MLOps" has come into the data lexicon as a combination of 1) Data Engineering, 2) Data Science, and 3) DevOps. A person in your organization may even carry the title or role of MLOps as an engineer and leader with primary ownership of all three disciplines.

7. **How is data presented to consumers?**

User interfaces and dashboards are a critical part of data analytics, but they are not the only vehicle to view and consume data. Oftentimes, developers, analysts, and data scientists require access to data in a controlled and structured manner. An API endpoint attached to a data source or even a predictive machine learning model provides a way to share data with other constituents through appropriate authentication and authorization. API management services can attach additional features such as endpoint monitoring, throttling, and even monetization/chargeback for data access.

Furthermore, power users of your data assets may need raw data access in order to make their own transformations and analyze features that may seem uninteresting to other users and applications. Together with API access, exposing raw data in a controlled manner are necessary steps to scaling decisions and opening up the platform to additional use cases. The **types of data presentations are described below**.

Dashboard UI	Data API or Model API	Raw Data
On-platform	Off-platform/External	On-platform
Interactive data visualization filled with the outputs of known methods and formulas from data where you already know the questions you needed to ask.	Provides controlled and structured access to data through an endpoint often over HTTP/S. APIs provide the opportunity to monitor how consumers are using the API, govern access and potentially even...	Exposes direct access to data repositories for consumers that need to ask their own questions and solve for their own objectives all while using the same data. Providing direct access to the data allows additional...

	monetize the endpoint as a service. This could be data ultimately retrieved from a database or even a predictive model that a 3rd party application can call to make a decision.	consumers to create their own metrics, dashboards, and predictive models while scaling decisions for other use cases.

8. **What requirements are there for data governance?**

If your data have value (which it most certainly does if you are exploring a platform to build around it) then a governance model helps ensure how data are accessed, managed, monitored, consumed, and stored. There may be many reasons to govern data that can involve intellectual property, laws and regulations, compliance, data valuation, privacy, and many others.

Just as it is important to enumerate data formats and the questions you are asking from your data, it is critical to develop and continually review a plan for governing data that will involve appropriate security controls, workforce training, and policy documentation.

Software solutions and cloud services offer helpful functions for governing data, but these are not the only considerations when developing a data governance strategy. Assigning ownership and evaluating data governance plans are necessary for understanding how to best utilize your data responsibly.

9. **What is the distance between the data source and the consumer(s)?**

For the purposes of this paper, distance involves the time and resources required to ingest data from the original source into a consumable format for users. In nearly all circumstances, the shortest distance is the most ideal. In a production environment, there is no such thing as zero distance because what might be ideal for users is not ideal or realistic for the enterprise. Therefore, we add on distance for reasons like governance, processing, security, etc.

What is important here is that the tradeoff between distance and production usability is made deliberately and consciously. For instance, simply adding a process without understanding the tradeoff increases the distance and makes it more difficult to justify and measure necessary added processing in the future.

Provides flexibility for uncovering new and potentially unforeseen use cases derived from access to new data or the blending of data sources, unlocking a diverse set of possibilities and outcomes.

1. **Have there been any insights, questions answered, or decisions made that should influence further aggregation from existing and new data sources?**

Data aggregation and storage can become increasingly expensive especially if you are not tracking the Return on Investment (ROI) metrics, or worse, the data you are actively collecting and processing are being ignored. Nevertheless, new data presents new opportunities through new discoveries or increasing the accuracy of existing discoveries. Finding the balance can sometimes be nontrivial, but it lies in measuring the cost of decisions you would like to make. Costs can include financial investments, people resources, technology resources, etc.

Equally as important as measuring the cost is measuring the output of a decision and/or future decisions. Accuracy, confidence level, business/mission priority, timeliness, and method type are all examples of measurable labels to attach to a decision that can facilitate the value and need of acquiring data from new sources.

2. **Can decisions, methods, and/or data be monetized?**

With a data platform providing value to decision-makers you can consider if 3rd party partners or customers would take advantage of insights, platform services (e.g. predictive model in production or BI dashboard), or direct access to the data. In some cases, 3rd parties may bring their own data to expose to the rest of your ecosystem of partners and customers, enabling a revenue stream based on syndication. No matter how resources or services can be monetized, a monitoring framework (or chargeback model) based on usage statistics allows you to track value over time, the methods you create to make decisions, and the decisions themselves.

3. **Should the platform graduate and tune usage of additional methods?**

Machine learning can lead to new use cases, new methods for making decisions, and also even new questions to ask. The more that machine learning is adopted by new tenants and the more success you have with making predictions or classifications, the more likely the platform could become a factory of new and initially unforeseen analytic methods.

New technologies and cloud services like pre-trained model types are democratizing machine learning for consumers beyond experienced data scientists, which makes advanced methods a reality for enterprise data platforms. The rest of this paper is going to examine a hypothetical example of a data platform based on the realities of business requirements from large data sets.

A Generalized Reference Architecture

We will define a fictitious data platform reference architecture called “Insights Central” or “IC,” which is a shareable set of components and data-driven functions that lead analysts, developers, and business leaders toward better decisions from data. While the platform started out as a traditional data warehouse, the executive stakeholders are looking to IC to provide insights from machine learning using predictive and classification models deployed into production and backed by APIs. **IC will provide:**

1. Data integration tools to acquire, aggregate, and transform data from multiple sources into common, centralized repositories.
2. A way to visualize data through both pre-built and newly developed dashboards in addition to a querying engine for viewing and manipulating structured and semi-structured data.
3. Machine learning pipelines that allow for orchestration of the model development and delivery lifecycle that includes data ingestion, feature engineering, model training, and model serving.
4. Digital platform APIs to access workflows, frameworks, and execution to deliver and publish customer applications that provide end-user access to data and trained models.
5. An API management framework that allows for developers and data scientists to build services and models utilizing Platform APIs that provide access to raw data, other trained models, dashboards, etc.

The **architectural characteristics for each component in IC** along with the **degree to which each characteristic is purposed** are described in the table below.

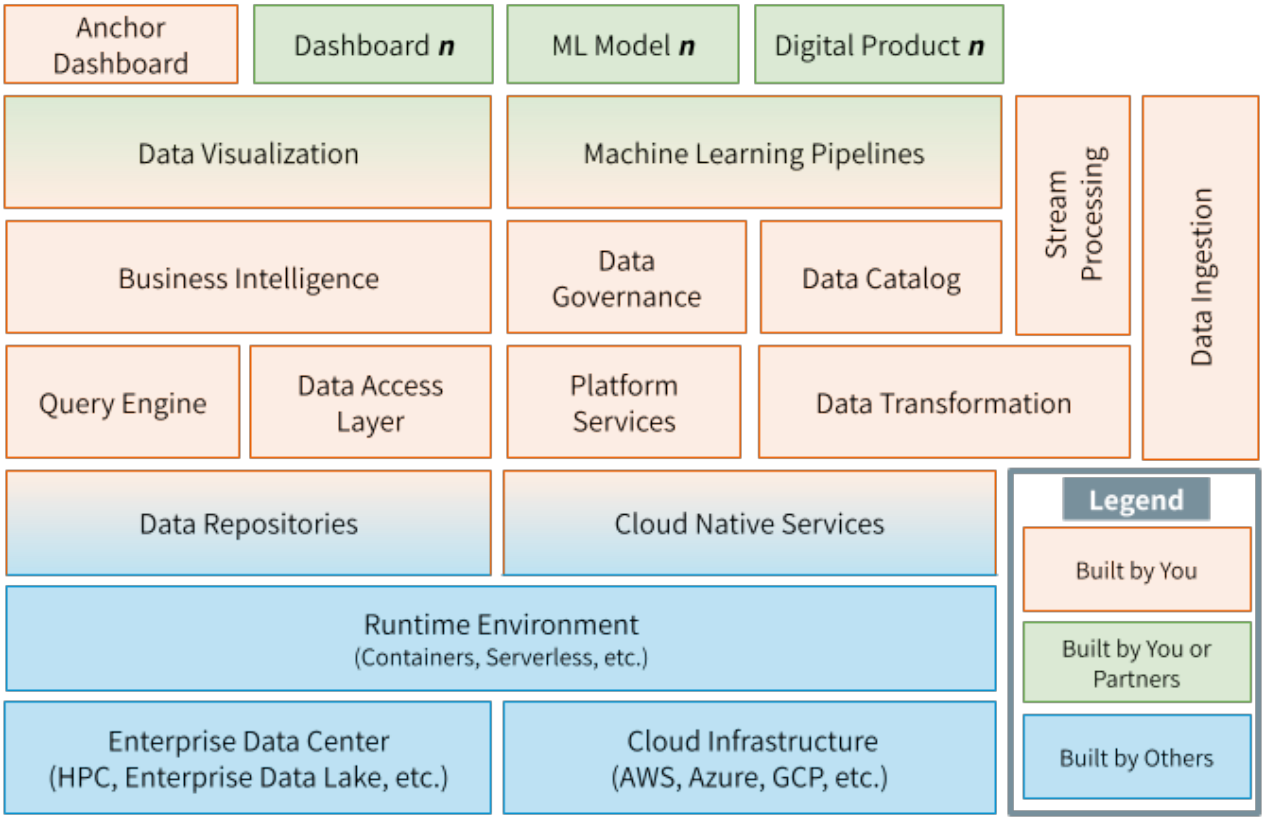
Architectural Characteristic	Degree	Description
Objective	Democratization of data	IC provides access to a petabyte-scale (PB) data warehouse with a host of tools for transformations and analysis. The IC platform is available for citizen users to access data, utilize known methods, and create new methods. The business wants to track and measure which methods are providing the most value to mission outcomes.
Method	Known and Unknown	IC initially stored data in data warehouses for BI use cases, but business stakeholders want to explore new methods through the use of machine learning.
Acquisition	Consolidation, Propagation through Event Stream Processing	Data arrive into IC repositories through multiple avenues, which is important to acknowledge the difference in formats and requirements for preparation. Such avenues include batch ETL and both synchronous and asynchronous data publishing to event streams.
Quality	Duplicated, Inaccurate, Inconsistent, and Missing Data	Low-quality data commonly show up in IC data pipelines depending on the source and time of acquisition. As new data types are explored for new insights, data transformation development and execution becomes critical to achieve and maintain high data quality.
Diversity	JSON, SQL, NoSQL, Event Streams, and Messages	The variety of applications that publish new data types and the emergence of new IoT devices from multiple vendors is what makes the data so varied between JSON, CSV, XML, SQL, NoSQL, et al.).

Consumers	Variety	Engineering, architecture, data science, business, and everything in between represents a range of stakeholders interested in the outcomes of IC.
Visualization	BI-based Dashboards, SQL, Data Preparation Engine	Dashboards represent the majority of data visualizations because they help a multitude of business stakeholders both internally and from 3rd Parties (Including executive-level sponsors).
Governance	Moderately Regulated	IC holds mostly Personal Identifiable Information (PII) from many geographies and coming up with a global governance strategy is nontrivial. In many cases, data platforms utilize repeatable data masking and encryption services that allow for the protection of sensitive data types or attributes.
Data Distance	No concept of measuring tradeoffs for additional data processing	Data engineering and governance teams add processing steps to IC at will based on continuous user, business, and compliance demands. Measuring the value of data sources, insights, and tradeoffs for adding distance through data catalogs and monitoring services will help justify the necessity of additional processing steps.
Further Data Acquisition	Edge, IoT	With the launch of edge-based IoT devices, IC is able to collect far more data compared to previous years. This will only increase as edge devices and software enhancements increase as well.
Monetization	Internal/Chargeback	While IC is sharing data and methods with 3rd Party partners, it is difficult for the organization to develop a monetization strategy. IC platform owners are able to provide a chargeback model to internal users who plan to utilize platform services to pursue their own data objectives.
Method Tuning	BI, AI/ML/DL, AR	While the BI practice is more established, and IC is looking to make machine learning more repeatable, Augmented Reality (AR) and Virtual Reality (VR) serve as areas to investigate for new acquisition strategies of data as well as new digital products that make use of emerging data services.

IC needs to include the following components in order to accelerate decisions using the organization's best-known methods while facilitating the development of new methods.

The journey to a data platform needs to achieve incremental milestones that must include data-driven outcomes. While the following components can be considered an ideal "desired state," initial data analysis or

integration should be labeled as the "Anchor," which can be anything that facilitates discovery (e.g. dashboard, report, custom UI, or API endpoint). The initial milestones that lead to outcomes that propel the business will trigger the appetite and investment for further platform development.



To better understand this logical architecture it is important to understand the role of each component, starting from the bottom of the diagram and working up. For each, a general description and a non-exhaustive list of technology options are provided below:

1. **Enterprise Data Center** - To date, IC has traditionally relied on High Performance Computing (HPC) as well as on-premises storage for their data. Due to historical data collection mechanisms and storage costs, many enterprise data platforms still retain some on-premises assets.

    a. **Technology Options**: Dell EMC/VMware, NetApp, Nutanix, Databricks, OpenStack, Ceph

    b. **Enabled Characteristics**: Hyper-Converged Resources, Infrastructure Orchestration

2. **Cloud Infrastructure** - IC is shifting to a cloud-based model for agility, access to new services, and scalability for data storage in addition to investing in cloud services that accelerate machine learning model training. Cloud computing, storage, and networking provide scalable resources without the

same management overhead as in on-premises environments.

a. **Technology Options**: AWS, Azure, GCP, and NVIDIA GPUs for model training

b. **Enabled Characteristics**: Elastic Provisioning, Scalability, Accelerated Processing

3. **Runtime Environment** - There are numerous reasons to make use of computing resources like virtual machines, containers, and container orchestration such as applications and APIs that surface data or trained models for predictions and classifications. Containers could also be used to run model training if there are customization requirements in machine learning libraries like TensorFlow or Pytorch that are not available in managed services (e.g. AWS Sagemaker, Google Vertex AI).

   a. **Technology Options**: Kubernetes, AWS EC2, Azure VM, Google Compute Engine

   b. **Enabled Characteristics**: Elastic Scalability, Embedded Runtimes, Customizable Configuration

4. **Data Repositories** - Can be one or many services that store data for analysis by a person or another service. Generally speaking, data becomes gradually more structured as it becomes used for analysis and decision-making.

   a. **Technology Options**:

      I. **Data Warehouse** - Structured storage for supporting BI-based analytics often in real-time (Google BigQuery, AWS Redshift, Databricks, Snowflake).

      II. **Data Lake** - Less structured, larger-scale, and longer-term retention of data that serves multiple use cases to include those not yet to-be-discovered (Google BigQuery, AWS S3. AWS Lake Formation, Azure Data Lake, Azure Synapse Databricks, Snowflake).

      NOTE: As of the publishing date of this paper, cloud providers are beginning to disambiguate between Data Warehouses and Data Lakes, matching queryable data for real-time analytics with PB-scale storage.

      III. **Event Streams** - Continuous delivery of data, typically in clustered computing environments. See the "Stream Processing" component below for a more detailed description.

      IV. **Database Management System (DBMS)** - Structured, transactional, and operational store most often utilized by applications and services. SQL and NoSQL are both viable options depending on data types, volume, and integration requirements (Microsoft SQL Server, Oracle Database, MySQL, PostgreSQL, MongoDB, Google Bigtable, AWS DynamoDB, Azure CosmosDB).

   b. **Enabled Characteristics**: Serverless Compute, Elastic Storage, Performance Optimized

5. **Cloud-Native Services** - IC requires several cloud services to support production workloads, software delivery, monitoring, and data management that include authentication, audit logging, monitoring, CDN, CI/CD, etc. In a data engineering context, pipeline or job alerts or statistics could be plugged into

a dedicated monitoring cloud monitoring service.

While enumerating these services is not the focus of this paper, it warrants mentioning their benefits in aggregate as critical components of cloud-based digital platforms.

6. **Data Ingestion** - Data are generated from a variety of sources and it is important to the business to collect from sources that could lead to meaningful decisions especially when combined with additional data. It is a reasonable expectation that the number of sources will increase over time as well as the diversity of data between sources.

    a. **Technology Options**: AWS Glue, Google Cloud Data Fusion, Google Cloud Pub/Sub, Azure Data Factory, Informatica, Talend

    b. **Enabled Characteristics**: Batch, Micro-Batch, and Stream-Based Acquisition

7. **Stream Processing** - Typically provides publishing and subscribing capabilities to event streams that can continuously deliver data points and scale-out to support large data sets. Performing analysis directly on the streams can lead to better data acquisitions and can help inform further transformations on the data. Event streams can also be considered a Data Repository especially when defined as a "source of truth" for data in the organization.

    a. **Technology Options**: Apache Kafka, Confluent, AWS Kinesis, Google Cloud Pub/Sub, Azure Service Bus, Informatica, Talend

    b. **Enabled Characteristics**: Event-driven, Real-time, Fault Tolerant, Elastic Scalability

8. **Data Transformation** -The component where batch and stream data (depending on the type of data ingestion) are converted into something more usable for storage, querying, and even analysis. Common methods within transformation that aim to generate high data quality can include data cleaning and data validation. In some cases, maximizing high data quality may involve validating data against an external service or repository to ensure accuracy (e.g. checking if a physical address is correct or the right data schema is being used, etc.)

    The transformation process consists of multiple steps that begin with discovering and profiling data to understand its structure and characteristics. Next, data are discovered to determine the desired output such as combining data with other data or converting data to a more usable format (i.e. data mapping). Finally, data engineers author code-based functions to execute against the data to achieve the desired output. Functions can expand into orchestrating complex workflows across multiple pipelines or jobs that all may depend on one another resulting in Directed Acyclic Graphs (DAGs).

    a. **Technology Options**: Google Dataflow, Google Cloud Data Fusion, Google Cloud Composer, AWS Data Pipeline, AWS Glue, Azure Data Factory, Apache Airflow, Apache Spark, Informatica, Talend

    b. **Enabled Characteristics**: Historical or Near Real-time, Discoverable, Interactive

9. **Query Engine** - A SQL/NoSQL solution that provides a way to reason over data, oftentimes from multiple, federated sources that have no preexisting engine (e.g. data lake or cloud storage service). An engine allows for queries to be simplified and more unified for investigating, structuring, and making additional transformations on data.

10. **Data Access Layer** - An API for creating, inserting, updating, and deleting structured, transactional, and operational data via standard connectivity protocols and approaches. Mostly utilized by developers to perform operations on data in their applications. Most database management systems and data warehouses already expose data access layer interfaces, but it is important to describe all mechanisms for data access and connectivity in the organization for discoverability purposes but also to build access controls for proper governance.

    a. **Technology Options**: JDBC, ODBC, GraphQL, REST API

    b. **Enabled Characteristics**: Compatible across the platform and with 3rd party tooling

11. **Data Governance** - A framework, set of policies, and technology services that assess and enforce data masking, security, risk, and compliance over all data types, access controls, and data handling protocols within the data platform.

    a. **Technology Options**: Google Cloud Data Fusion, Google Cloud Data Loss Prevention AWS Glue, Azure Purview

    b. **Enabled Characteristics**: Data Discoverability, Data Lineage, Metadata Management, Data Auditing

12. **Data Catalog** - Metadata management for discovering, organizing, and scaling data assets. Data Catalog services are helpful during the analytics lifecycle and also play a key role in governance, monitoring, or auditing analytic or pipeline jobs.

    a. **Technology Options**: AWS Glue, Google Data Catalog, Azure Data Catalog

    b. **Enabled Characteristics**: Data Discoverability, Data Lineage, Metadata Management, Data Loss Prevention, Supplemental Pipeline Monitoring

13. **Digital Platform Services** - Associated shareable services required for any digital platform. Services such as tenant management, entitlements, and access to the ecosystem of APIs are what assist in graduating from a digital product to a digital platform.

14. **Data Visualization/Business Intelligence** - Workflow that consists of the acquisition and analysis of data based on known methods and business metrics sometimes referred to as KPIs. BI involves data mining, performance benchmarking, and descriptive analytics that can produce dashboards, reporting, performance measures, and trends that inform decisions.

    a. **Technology Options**: AWS QuickSight, Google Cloud Data Studio, Looker, Azure Power BI, Qlik, Tableau

b. **Enabled Characteristics:** Visual, Digestible, Collaborative

15. **ML Pipelines** - A phased, continuous process of gathering and preparing data for machine learning, extracting features, training models, and serving them into production. Machine learning requires extra, additional stages of data preparation and validation beyond those operations performed during data engineering pipelines.

6.

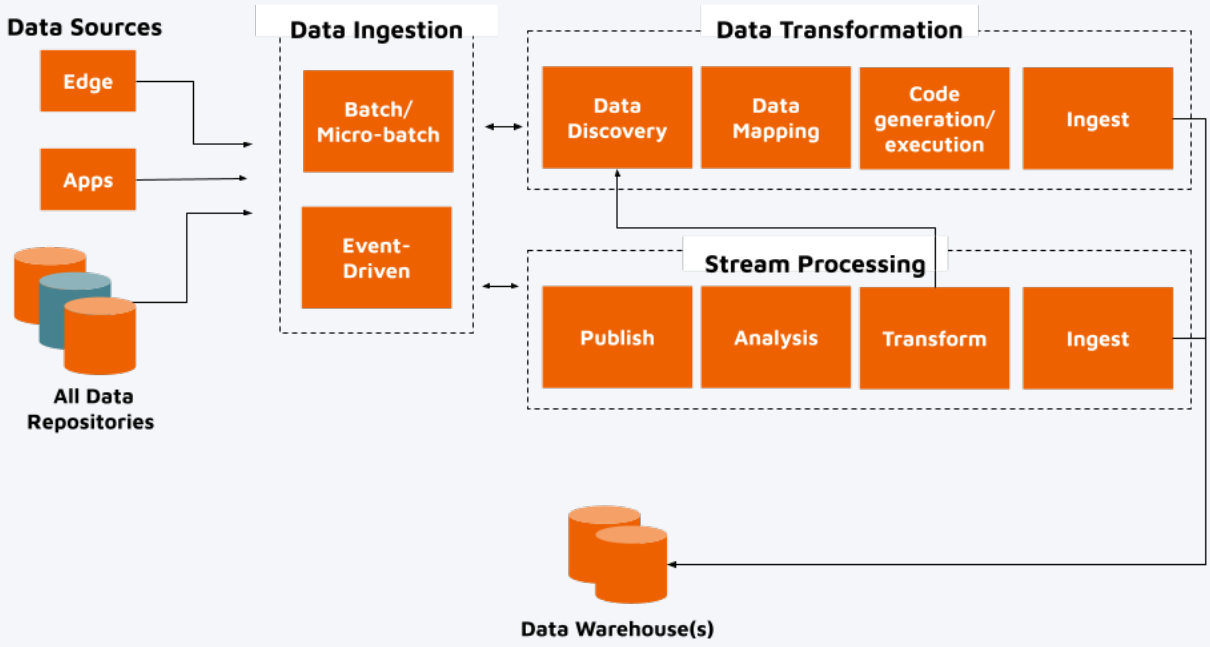
a. **Technology Options:** Google Vertex AI, Google AutoML, TensorFlow Extended, AWS Sagemaker, Azure Machine Learning, Databricks, IBM Watson

b. **Enabled Characteristics**: Elastic Scalability, Model Training, Model Evaluation, Pre-trained Models

Multiple services play critical roles in different stages of the data lifecycle. For the purposes of clarity, we **compiled all of the components above into a workflow and architecture** that is divided into two primary stages: **1)** Data Engineering and **2)** Data Analytics.

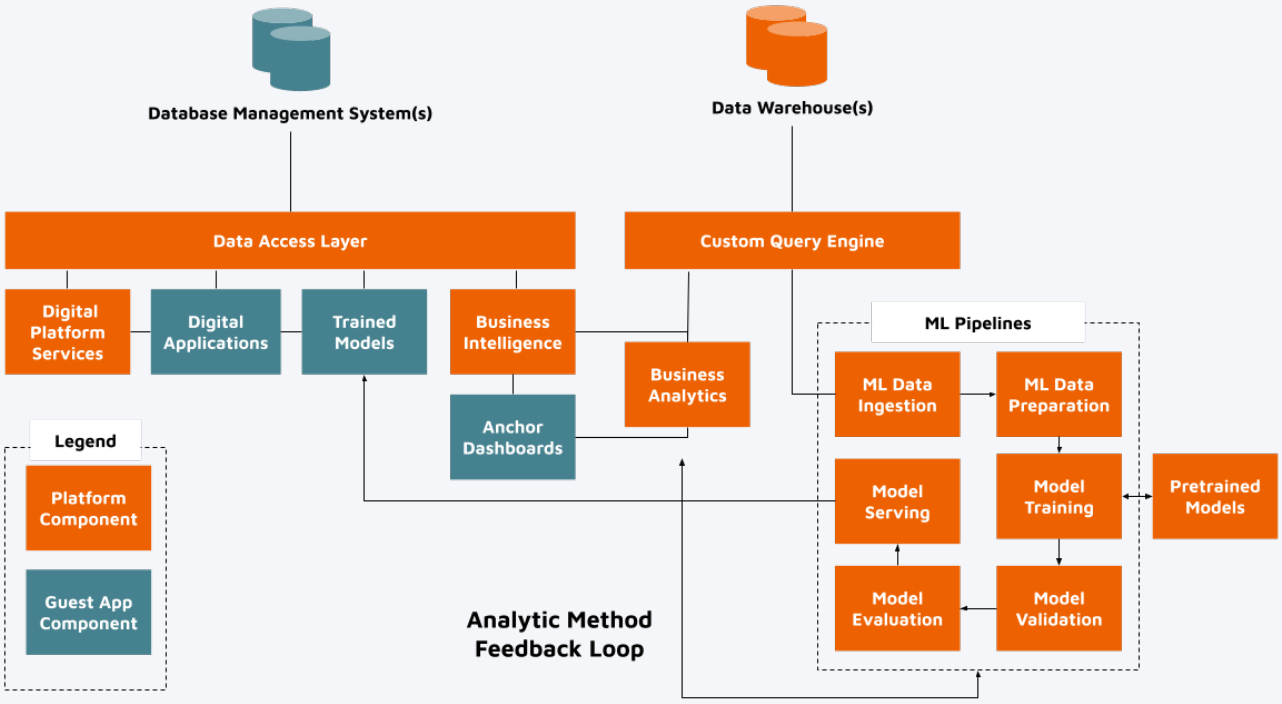
DATA ENGINEERING

As data are acquired, each source and format requires a targeted processing mechanism for converting that data into something usable. A data warehouse is a typical location for storing data that have already undergone initial processing via a data pipeline.



DATA ANALYTICS

Once processed, data can be queried for analysis and even further processing, especially when preparing data for Machine Learning. But even post-processed data are meaningless without it leading to a decision. Depending on the type of decision your organization requires, and the kinds of persona responsible for the decision, the data need to traverse through a workflow that will lead to a high-valued insight. Digital applications, trained models exposed as APIs, and BI dashboards are all options for providing insights on their own, but they can also be used collaboratively in feedback loops to inform other analytic workflows.



Conclusion

Setting goals for the decision-making outcomes in your organization is the first step to better understanding how to make the best use of your data. From there, you can begin to take stock in your sources, resources, and questions you would like to ask and answer, which can determine how you begin your platform journey. The intent of this paper was to give you the tools, the requirements, and some of the common components you will need to help scale decision-making in your organization.

ABOUT NUVALENCE

Nuvalence is a next-generation consulting firm specializing in digital platform and product development. Using our Product-Driven approach, we help organizations across all industries build impactful software solutions. Our software engineers, product managers, and designers find ways to accelerate innovation by leveraging our technical expertise and strong experience in commercial software engineering. We don't just deliver software, we deliver outcomes.