unravel™

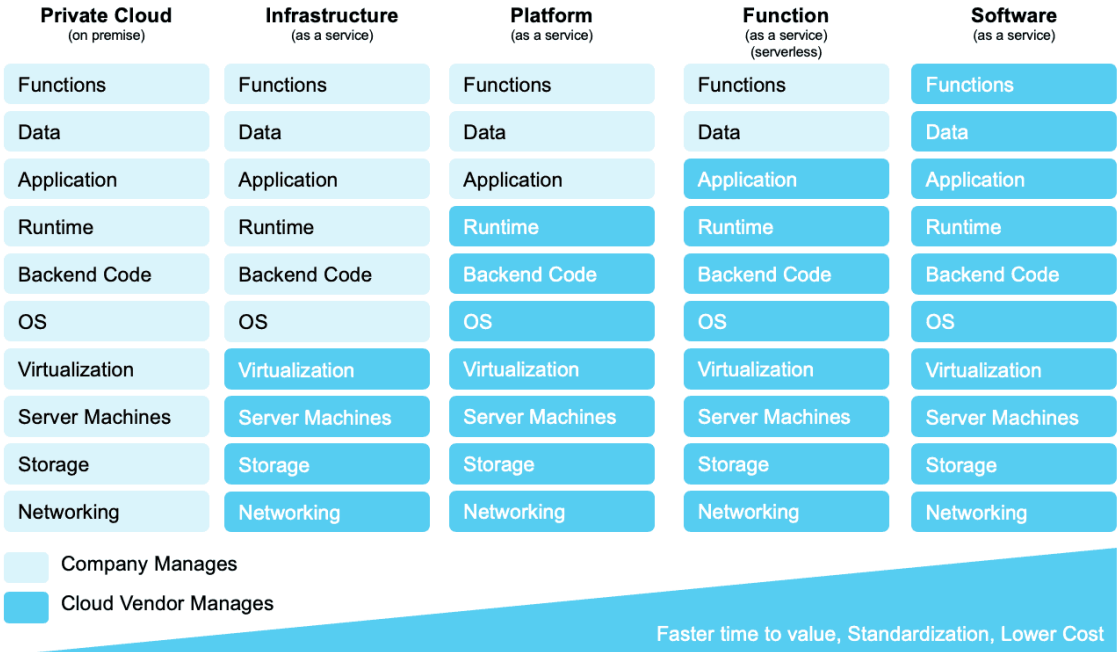
UNDERSTANDING CLOUD DATA SERVICES

Unravel Blog post by guest author, **Charlie Crocker, Analytics Systems and Product Management Leader**

Demystifying service offerings from Microsoft Azure, Amazon AWS and Google Cloud Platform

In the past five years, a shift in Cloud Vendor offerings has fundamentally changed how companies buy, deploy and run big data systems. Cloud Vendors have absorbed more back-end data storage and transformation technologies into their core offerings and are now highlighting their data pipeline, analysis, and modeling tools. This is great news for companies deploying, migrating, or upgrading big data systems. Companies can now focus on generating value from data and Machine Learning (ML), rather than building teams to support hardware, infrastructure, and application deployment/monitoring.

The following chart shows how more and more of the cloud platform stack is becoming the responsibility of the Cloud Vendors (shown in blue). The new value for companies working with big data is the maturation of Cloud Vendor Function as a Service (FaaS), also known as serverless, and Software as a Service (SaaS) offerings. For FaaS (serverless) the Cloud Vendor manages the applications and users focus on data and functions/features. With SaaS, features and data management become the Cloud Vendor's responsibility. Google Analytics, Workday, and Marketo are examples of SaaS offerings.



As the technology gets easier to deploy, and the Cloud Vendor data services mature, it becomes much easier to build data-centric applications and provide data and tools to the enterprise. This is good news: companies looking to migrate from on-premise systems to the cloud are no longer required to purchase directly or manage hardware, storage, networking, virtualization, applications, and databases. In addition, this changes the operational focus for a big data systems from infrastructure and application management (DevOps) to pipeline optimization and data governance (DataOps). The following table shows the different roles required to build and run Cloud Vendor-based big data systems.

On Premise	Infrastructure as a Service	Platform as a Service	Function as a Service	Software as a Service
Data Operations	Data Operations	Data Operations	Data Operations	Data Operations
Data Engineers	Data Engineers	Data Engineers	Data Engineers	Data Engineers
Pipeline Engineers	Pipeline Engineers	Pipeline Engineers	Pipeline Engineers	
Solution Engineers	Solution Engineers	Solution Engineers		
Infra Engineers	Infra Engineers			
HW/NW Engineers				

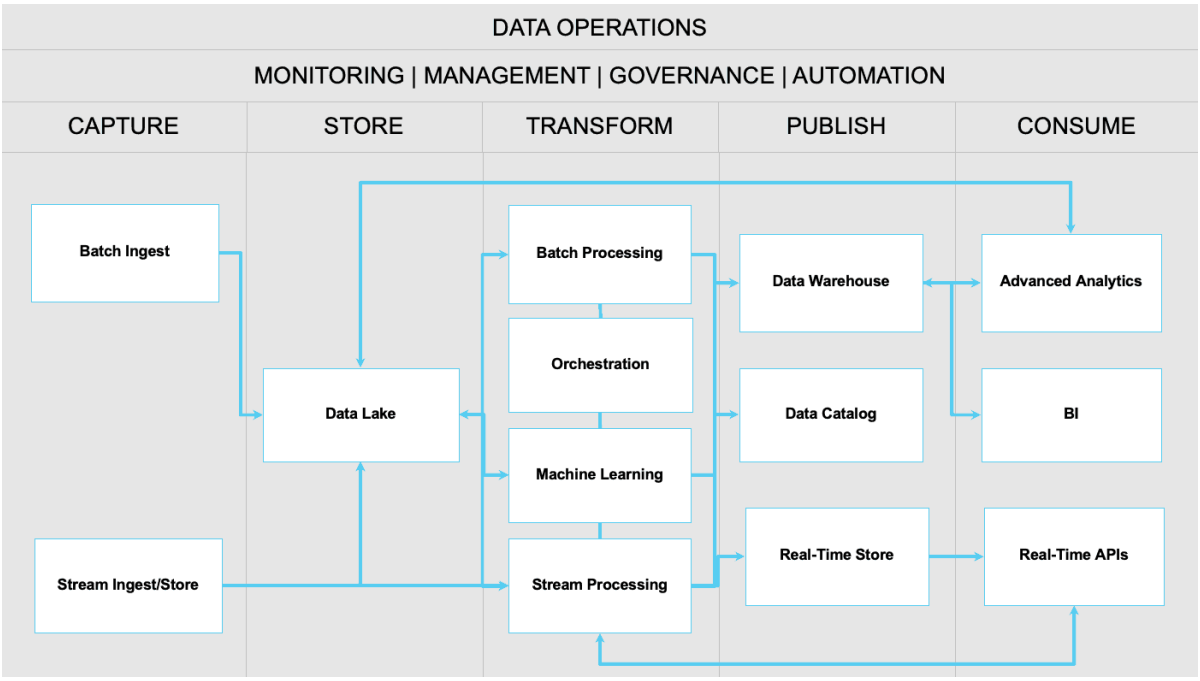
This article is aimed at helping big data systems leaders moving from on-premise or native IaaS (compute, storage, and networking) deployments understand the current Cloud Vendor offerings. Those readers new to big data, or Cloud Vendor services, will get a high-level understanding of big data system architecture, components, and offerings. To facilitate discussion we provide an end-to-end taxonomy for big data systems and show how the three leading Cloud Vendors (AWS, Azure and GCP) align to the model:

- Amazon Web Services (AWS)
- Microsoft Azure (Azure)
- Google Cloud Platform (GCP)

Applying a common taxonomy

Understanding Cloud Vendor offerings and big data systems can be very confusing. The same service may have multiple names across Cloud Vendors and, to complicate things even more, each Cloud Vendor has multiple services that provide similar functionality. However, the Cloud Vendors big data offerings align to a common architecture and set of workflows.

Each big data offering is set up to receive high volumes of data to be stored and processed for real-time and batch analytics as well as more complex ML/AI modeling. In order to provide clarity amidst the chaos, we provide a two-level taxonomy. The first-level includes five stages that sit between data sources and data consumers: CAPTURE, STORE, TRANSFORM, PUBLISH, and CONSUME. The second-level taxonomy includes multiple service offerings for each stage to provide a consistent language for aligning Cloud Vendor solutions.



The following sections provide details for each stage and the related service offerings.

Capture

Persistent and resilient data CAPTURE is the first step in any big data system. Cloud Vendors and the community also describe data CAPTURE as ingest, extract, collect, or more generally as data movement. Data CAPTURE includes ingestion of both batch and streaming data. Streaming event data becomes more valuable by being blended with transactional data from internal business applications like SAP, Siebel, Salesforce, and Marketo. Business application data usually resides within a proprietary data model and needs to be brought into the big data system as changes/transactions occur.

Cloud Vendors provide many tools for bringing large batches of data into their platforms. This includes database migration/replication, processing of transactional changes, and physical transfer devices when data volumes are too big to send efficiently over the internet. Batch data transfer is common for moving on-premise data sources and bringing in data from internal business applications, both SaaS and on-premise. Batch data can be run once as part of an application migration or in near real-time as transactional updates are made in business systems.

The focus of many big data Pipeline implementations is the capture of real-time data streaming in as an application clickstream, product usage events, application logs, and IoT sensor events. To properly capture streaming data requires configuration on the edge device or application. For, example, collecting clickstream from a mobile or web application requires events to be instrumented and sent back to an endpoint listening for the events. This is similar with IoT devices, which may also perform some data processing on the edge device prior to sending it back to an end point.

Store

For big data systems the STORE stage focuses on the concept of a data lake, a single location where structured, semi-structured, unstructured data and objects are stored together. The data lake is also a place to store the output from extract, transform, load (ETL) and ML pipelines running in the TRANSFORM stage. Vendors focus on scalability and resilience over read/write performance. To increase data access and analytics performance, data should be highly aggregated in the data lake or organized and placed into higher performance data warehouses, massively parallel processing (MPP) databases, or key-value stores as described in the PUBLISH stage. In addition, some data streams have such high event volume, or the data are only relevant at the time of capture, that the data stream may be processed without ever entering the data lake.

Cloud Vendors have recently put more focus on the concept of the data lake, by adding functionality to their object stores and creating a much tighter integration with TRANSFORM and CONSUME service offerings. For example, Azure created Data Lake Storage on top of the existing Object Store with additional services for end to end analytics pipelines. Also, AWS now provides Data Lake Formation to make it easier to set up a data lake on their core object store S3.

Transform

The heart of any big data implementation is the ability to create data pipelines in order to clean, prepare, and TRANSFORM complex multi-modal data into valuable information. Data TRANSFORM is also described as preparing, massaging, processing, organizing, and analyzing among other things. The TRANSFORM stage is where value is created and, as a result, Cloud Vendors, start-ups, and traditional database and ETL vendors provide many tools. The TRANSFORM stage has three main data pipeline offerings including Batch Processing, Machine Learning, and Stream Processing. In addition, we include the Orchestration offering because complex data pipelines require tools to stage, schedule, and monitor deployments.

Batch TRANSFORM uses traditional extract, TRANSFORM, and load techniques that have been around for decades and are the purview of traditional RDBMS and ETL vendors. However, with the increase in data volumes and velocity, TRANSFORM now commonly comes after extraction and loading into the data lake. This is referred to as extract, load, and transform or ELT. Batch TRANSFORM uses Apache Spark or Hadoop to distribute compute across multiple nodes to process and aggregate large volumes of data.

ML/AI uses many of the same Batch Processing tools and techniques for data preparation and for the development and training of predictive models. Machine Learning also takes advantage numerous libraries and packages to help optimize data science workflows and provide pre-built algorithms.

big data systems also provide tools to query continuous data streams in near real-time. Some data has immediate value that would be lost waiting for a batch process to run. For example, predictive models for fraud detection or alerts based on data from an IoT sensor. In addition, streaming data is commonly processed, and portions are loaded into a data lake.

Cloud Vendor offerings for TRANSFORM are evolving quickly and it can be difficult to understand which tools to use. All three Cloud Vendors have versions of Spark/Hadoop that scale on their IaaS compute nodes. However, all three now provide serverless offerings that make it much simpler to build and deploy data pipelines for batch, ML and streaming workflows. For example, AWS EMR, GCP Cloud Data Proc, and Azure Databricks provide Spark/Hadoop that scale by adding additional compute resources. However, they also offer the serverless AWS Glue, GCP Data Flow, and Azure Data Factory which abstract away the need to manage compute nodes and orchestration tools. In addition, they now all provide end-to-end tools to build, train, and deploy machine learning models quickly. This includes data preparation, algorithm development, model training algorithm, and deployment tuning and optimization.

Publish

Once through the data CAPTURE and TRANSFORM stages it is necessary to PUBLISH the output from batch, ML, or streaming pipelines for users and applications to CONSUME. PUBLISH is also described as deliver or serve, and comes in the form of Data Warehouses, Data Catalogs, or Real-Time Stores.

Data Warehouse solutions are abundant in the market, and the choice depends on the data scale and complexity as well as performance requirements. Serverless relational databases are a common choice for Business Intelligence applications and for publishing data for other systems to consume. They provide scale and performance and, most of all, SQL-based access to the prepared data. Cloud Vendor examples include AWS Redshift, Google BigQuery, and Azure SQL Data Warehouse. These work great for moderately sized and relatively simple data structures. For higher performance and complex relational data models, massively parallel processing (MPP) databases store large volumes of data in-memory and can be blazing fast, but often at a steep price.

As the tools for TRANSFORM and CONSUME become easier to use, data analyses, models, and metrics proliferate. It becomes harder to find valuable, governed, and standardized metrics in the mass of derived tables and analyses. A well-managed and up-to-date data catalog is necessary for both technical and non-technical users to manage and explore published tables and metrics. Cloud Vendor Data Catalog offerings are still relatively immature. Many companies build their own or use third party catalogs like Alation or Waterline. More technical users including data engineers and data scientists explore both raw and transformed data directly in the data lake. For these users the data catalog, or metastore, is the key for various compute options to understand where data is and how it is structured.

Many streaming applications require a Real-Time Store to meet millisecond response times. Hundreds of optimized data stores exist in the market. As with Data Warehouse solutions, picking a Real-Time Store depends on the type and complexity of the application and data. Cloud Vendors examples include AWS DynamoDB, Google Bigtable, and Azure Cosmos DB providing wide-column or key-value data stores. These are applied as high performance in-process databases and improve the performance of Data processing and analytics workloads.

Consume

The value of any big data system comes together in the hands of technical and non-technical users, and in the hands of customers using data-centric applications and products. Vendors also refer to CONSUME as use, harness, explore, model, infuse, and sandbox. We discuss three CONSUME models: Advanced Analytics, Business Intelligence (BI), and Real-Time APIs.

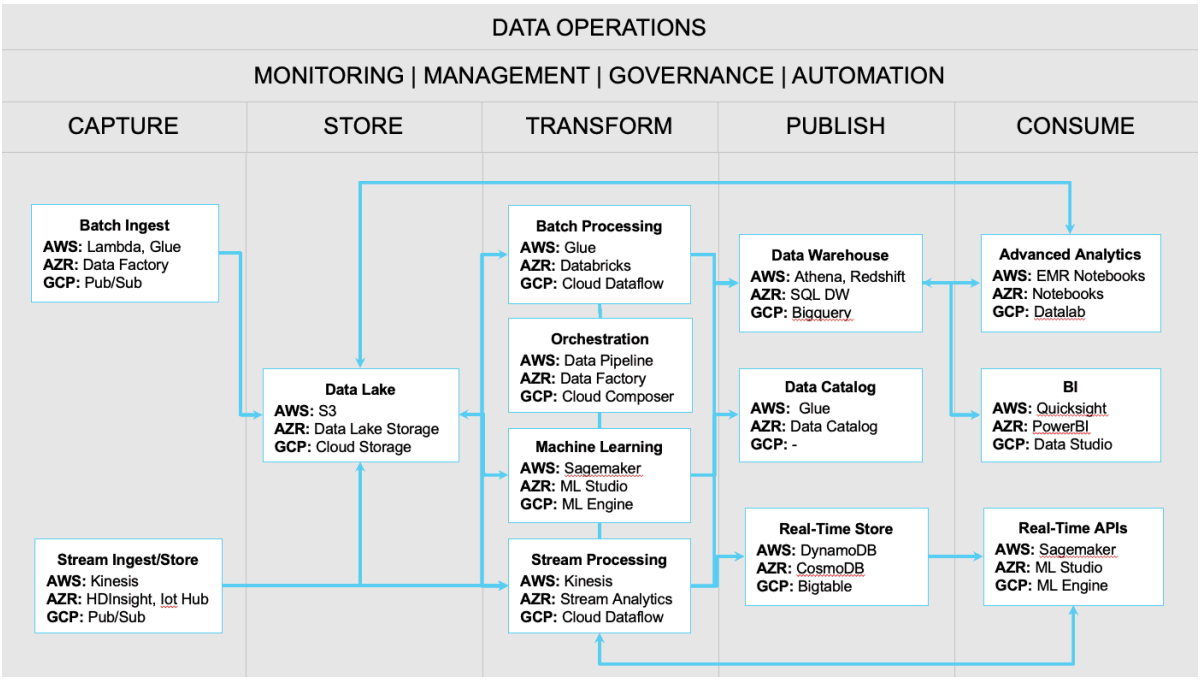
Aggregated data does not always allow for deeper data exploration and understanding. So, advanced analytics users CONSUME both raw and processed data either directly from the data lake or from a Data Warehouse. Advanced analytics users use similar tools from the TRANSFORM stage including Spark- and Hadoop-based distributed compute. In addition, notebook technologies are a popular tool that allow data engineers and data scientists to create documents containing live code, equations, visualizations and text. Notebooks allow users to code in a variety of languages, run packages, and share the results. All three Cloud Vendors offer notebook solutions, most based on the popular open source Jupyter project.

BI tools have been in the market for a couple of decades and are now being optimized to work with larger data sets, new types of compute, and directly in the cloud. Each of the three cloud vendors now provide a BI tool optimized to work with their stack. These include AWS Quicksight, GCP Data Studio, and Microsoft Power BI. However, several more mature BI tools exist in the market that work with data from most vendors. BI tools are optimized to work with published data and usage improves greatly with an up-to-date data catalog and some standardization of tables and metrics.

Applications, products, and services also CONSUME raw and transformed data through APIs built on the Real-Time Store or predictive ML models. The same Cloud Vendor ML offerings used to explore and build models also provides Real-Time APIs for alerting, analysis, and personalization. Example use cases include fraud detection, system/sensor alerting, user classification, and product personalization.

Cloud vendor offerings

AWS, GCP and AZURE have very complex cloud offerings based on their core networking, storage and compute. In addition, they provide vertical offerings for many markets, and within the big data systems and ML/AI verticals they each provide multiple offerings. In the following chart we align the Cloud Vendor offerings within the two-tier big data system taxonomy defined in the second section.



The following table includes some additional Cloud Vendor offerings as well as open source and selected third party tools that provide similar functionality.

	AWS	Azure	GCP	Open Source	3rd Party
CAPTURE					
Batch ingest	Amazon Kinesis Lambda	Data Factory	Cloud Pub/Sub	Sqoop Nifi	Attunity GoldenGate
Stream ingest	nazon Managed Streaming for Kaf Kinesis Data Streams Kinesis Data Firehose	Event Hubs IoT Hub	Cloud Pub/Sub	Kafka Flume	Amplitude Segment
STORE					
Datalake	AWS Lake Formation mazon Simple Storage Service (S3)	Azure Data Lake Storage	Cloud Storage	HDFS	
TRANSFORM					
Batch Data Processing	Amazon EMR AWS Glue AWS Data Pipeline AWS Batch	Azure Databricks Data Lake Analytics Azure Analysis Services HDInsight	Cloud Dataprep Cloud Dataproc Cloud Dataflow	Nifi Spark/Hive	Matillion Talend
Orchestration	Data Pipeline	Data Factory HDInsight	Cloud Composer		
Machine Learning	SageMaker	Machine Learning Studio R Server for HDInsight	Machine Learning Engine	Mahout TensorFlow	Dataiku Data Robot
Stream Processing	Kinesis Data Analytics	Azure Data Explorer Azure Stream Analytics HDInsight Data Factory	Cloud Dataflow	Storm Spark	Confluent Streamsets
PUBLISH					
Analytics Warehouse	Amazon Athena Amazon Redshift	SQL Data Warehouse	BigQuery	MariaDB Druid	Snowflake Exasol
Catalog-Metastore	AWS Glue	Data Catalog		HCatelog Hive	Alation Waterline
Real-Time Store	Amazon DynamoDB	Azure Cosmos DB	Cloud Bigtable	Casandra Hbase	MongoDB MarkLogic
CONSUME					
Business Intelligence	Amazon QuickSight	Power BI Embedded	Cloud Datalab	BIRT Superset	Tableau Looker
Advanced Analytics	EMR Notebooks	Notebooks	Google Data Studio	Zeppelin	Talend Databricks
Real-Time APIs	SageMaker	Machine Learning Studio	Machine Learning Engine	H2O	Dataiku Data Robot

The time is now

Most companies that deployed big data systems and data-centric applications in the past 5-10 years did this on-premise (or colocation) or on top of the Cloud Vendor core infrastructure services including storage, networking, and compute. Much has changed in the Cloud Vendor offerings since these early deployments. Cloud Vendors now provide a nearly complete set of serverless big data services. In addition, more and more companies see the value of Cloud Vendor offerings and are trusting their mission-critical data and applications to run on them. So, now is the time to think about migrating big data applications from on-premise or upgrading bespoke systems built on Cloud Vendor infrastructure services. In order to make the best use of, make sure to get a deep understanding of existing systems, develop a clear migration strategy, and establish a data operations center of excellence.

In order to prepare for migration to Cloud Vendor big data offerings, it is necessary for an organization to get a clear picture of its current big data system. This can be difficult depending on the heterogeneity of existing systems, the types of data-centric products it supports, and the number of teams or people using the system. Fortunately, tools (such as Unravel) exist to monitor, optimize, and plan migrations for big data systems and pipelines. During migration planning it is common to discover inefficient, redundant, and even unnecessary pipelines actively running, chewing up compute, and costing the organization time and money. So, during the development of a migration strategy companies commonly find ways to clean up and optimize their data pipelines and overall data architecture.

It is helpful that all three Cloud Vendors are interested in getting a company's data and applications onto their platforms. To this end, they provide a variety of tools and services to help move data or lift and shift applications and databases onto their platforms. For example, AWS provides a Migration Hub to help plan and execute migrations and a variety of tools like the AWS Database Migration Service. Azure provides free Migration Assessments as well as several tools. And, GCP provides a variety of migration strategies and tools like Anthos and Velostrata depending on a company's' current and future system requirements.

Please take a look at the Cloud Vendor migration support sites below.

- Google Cloud Platform (GCP)
- Microsoft Azure
- Amazon AWS

No matter whether a company runs an on-premise system or a fully managed serverless environment, or some hybrid combination, companies need to build expertise a core competence in data operations. DataOps is a rapidly emerging discipline that companies need to own–it is difficult to outsource. Most data implementations utilize tools from multiple vendors, maintain hybrid cloud/on-premises systems, or rely on more than one Cloud Vendor. So, it becomes difficult to rely on a single company or Cloud Vendor to manage all the DataOps task for an organization.

Typical scope includes:

- Data quality
- Metadata management
- Pipeline optimization
- Cost management and Charge back
- Performance Management
- Resource Management
- Business stakeholder Management
- Data governance
- Data catalogs
- Data security & compliance
- ML/AI model management
- Corporate metrics and reporting

Where ever you are on your cloud adoption and workload migration journey, now is the time to start or accelerate your strategic thinking and execution planning for Cloud based data services. Serverless offerings are maturing quickly and give companies faster time to value, increased standardization, and overall lower people and technology costs. However, as migration goes from planning to reality, ensure you invest in the critical skills, technology and process changes to establish a data operations center of excellence.

Reach out to us to start your **free, full-featured trial** at **hello@unraveldata.com**

About Unravel

Unravel radically simplifies the way businesses understand and optimize the performance of their modern data applications – and the complex pipelines that power those applications. Providing a unified view across the entire stack, Unravel's AI-powered data operations platform leverages AI, machine learning, and advanced analytics to offer actionable recommendations and automation for tuning, troubleshooting, and improving performance – both today and tomorrow. By operationalizing how you do data, Unravel's solutions support modern big data leaders, including Kaiser Permanente, TIAA, Adobe, Deutsche Bank, Wayfair, and Neustar. The company is headquartered in Palo Alto, California, and is backed by Menlo Ventures, GGV Capital, M12, Data Elite Ventures, and Jyoti Bansal. To learn more, visit unraveldata.com.

unravel

unraveldata.com | hello@unraveldata.com

Reach out to us to start your **free, full-featured trial** at **hello@unraveldata.com**

About Unravel

Unravel radically simplifies the way businesses understand and optimize the performance of their modern data applications – and the complex pipelines that power those applications. Providing a unified view across the entire stack, Unravel's AI-powered data operations platform leverages AI, machine learning, and advanced analytics to offer actionable recommendations and automation for tuning, troubleshooting, and improving performance – both today and tomorrow. By operationalizing how you do data, Unravel's solutions support modern big data leaders, including Kaiser Permanente, TIAA, Adobe, Deutsche Bank, Wayfair, and Neustar. The company is headquartered in Palo Alto, California, and is backed by Menlo Ventures, GGV Capital, M12, Data Elite Ventures, and Jyoti Bansal. To learn more, visit unraveldata.com.

unravel

unraveldata.com | hello@unraveldata.com