privacera

WHITEPAPER

Rethinking the Modern Data Stack for the Age of Gen AI

privacera

contents

Origins of Modern Data Stack Thinking

The term Modern Data Stack was coined as cloud data warehouses began gaining prominence. Without question, the cloud data solutions powered innovation by delivering specialized compute or processing services—essentially creating a framework that future-proofed architectural constructs. At the same time, cloud data warehouses demanded new approaches to data management, especially more interoperability for data ecosystems. With the complexities of a modern data ecosystem, using one vendor for everything was no longer an option.

For this reason, the Modern Data Stack refactored data operations in a way that proved more efficient and effective than legacy, on-premises data warehouses. It also drove federation of data ownership and stewardship. This moved organizations from centralized command and control to a federated approach with centralized oversight, observability, and more.

But in complete disclosure, the Modern Data Stack was initially conceived narrowly as a means of comparing on premises extract, transform, and load (ETL) versus more modern data ingestion approaches. With on-premises ETL, organizations extracted data from sources, transforming it, and then putting it into a data warehouse. This approach supported analytics that were mostly backwards-looking and financial in nature. For this reason, these data warehouses had carefully and rigidly designed models to support a select few analytical use cases.

The modern world supports the traditional use cases as well as forward-looking and predictive use cases. Here, designs and models are constructed on the fly and are often ephemeral. With the Modern Data Stack, organizations instead extract data directly in raw formats, including normalized data from transactional system sources. In these formats, data is loaded directly into the cloud. This change makes the load function more automated, moving the task of transformations and data modeling to the data lake or warehouse.

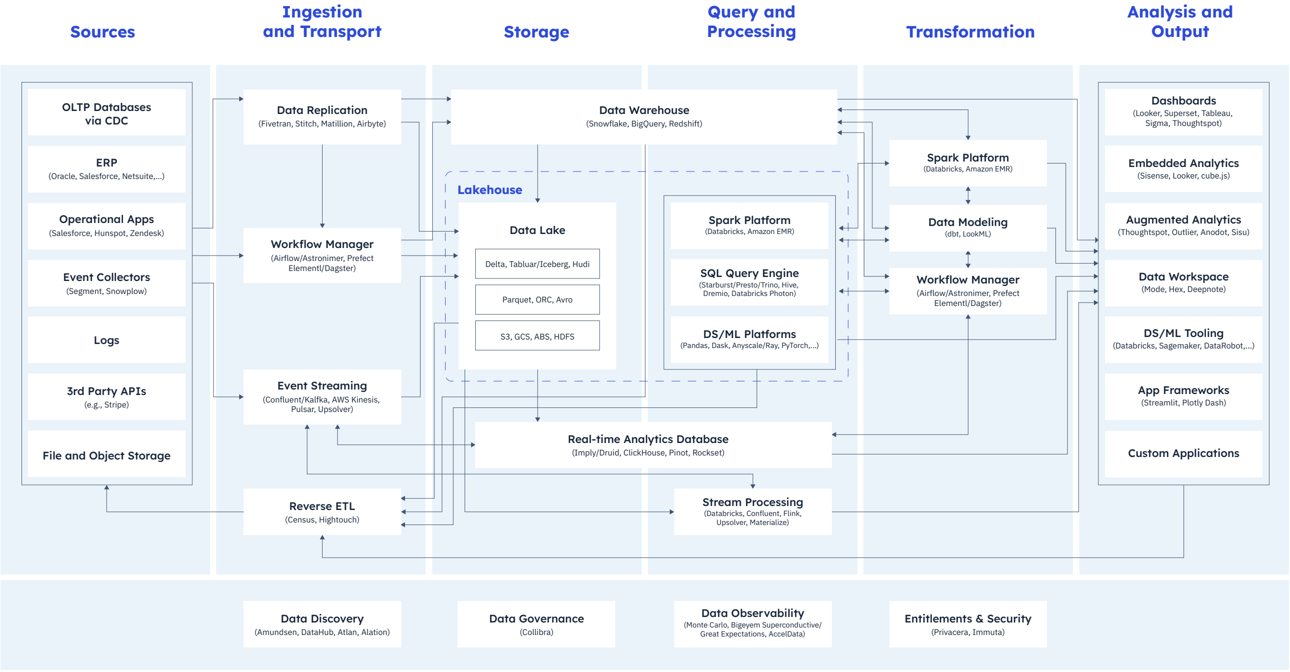
More recently, Matt Bornstein, Jennifer Li, and Martin Casado of Andreessen Horowitz published “Emerging Architectures for Modern Data Infrastructure.” This whitepaper expanded the definition of a Modern Data Stack and effectively defined the components of what major market analyst firms defined as a data fabric.The authors of “Rewired” say, “A data fabric is a modernized, centralized approach to data. What distinguishes the data fabric is the promise of greatly accelerated and cheaper integration through virtualization to connect data sources to the data fabric without unnecessary data movement.”

Put simply, Bornstein, Li, and Casado suggested a Modern Data Stack is about end-to-end systems, including sources, ingestion and transport, storage, query and processing, transformation, and analysis and output. Bornstein, Li, and Casadoa also added data discovery, data governance, data observability, and entitlements and security. Today, the Modern Data Stack once more needs updating to reflect the emergence of Generative AI and how each of the above functions support it.

What is a Modern Data Stack?

A Modern Data Stack manages the flow of data from raw data moving through ingestion and transport services into core data platforms that manage storage, query and processing, and transformation prior to being consumed by users in a variety of analysis and output modalities.

Importantly, it makes explicit the importance to data architecture of critical services, including data discovery, data governance, data observability, entitlements, and security.


Sources
Ingestion and Transport
Storage
Query and Processing
Transformation
Analysis and Output
OLTP Databases via CDC
ERP (Oracle, Salesforce, Netsuite,...)
Operational Apps (Salesforce, Hunspot, Zendesk)
Event Collectors (Segment, Snowplow)
Logs
3rd Party APIs (e.g., Stripe)
File and Object Storage
Data Replication (Fivetran, Stitch, Matillion, Airbyte)
Workflow Manager (Airflow/Astronimer, Prefect Elementl/Dagster)
Event Streaming (Confluent/Kalfka, AWS Kinesis, Pulsar, Upsolver)
Reverse ETL (Census, Hightouch)
Data Warehouse (Snowflake, BigQuery, Redshift)
Lakehouse
Data Lake
Delta, Tabluar/Iceberg, Hudi
Parquet, ORC, Avro
S3, GCS, ABS, HDFS
Spark Platform (Databricks, Amazon EMR)
SQL Query Engine (Starburst/Presto/Trino, Hive, Dremio, Databricks Photon)
DS/ML Platforms (Pandas, Dask, Anyscale/Ray, PyTorch,...)
Real-time Analytics Database (Imply/Druid, ClickHouse, Pinot, Rockset)
Stream Processing (Databricks, Confluent, Flink, Upsolver, Materialize)
Spark Platform (Databricks, Amazon EMR)
Data Modeling (dbt, LookML)
Workflow Manager (Airflow/Astronimer, Prefect Elementl/Dagster)
Dashboards (Looker, Superset, Tableau, Sigma, Thoughtspot)
Embedded Analytics (Sisense, Looker, cube.js)
Augmented Analytics (Thoughtspot, Outlier, Anodot, Sisu)
Data Workspace (Mode, Hex, Deepnote)
DS/ML Tooling (Databricks, Sagemaker, DataRobot,...)
App Frameworks (Streamlit, Plotly Dash)
Custom Applications
Data Discovery (Amundsen, DataHub, Atlan, Alation)
Data Governance (Collibra)
Data Observability (Monte Carlo, Bigeyem Superconductive/ Great Expectations, AccelData)
Entitlements & Security (Privacera, Immuta)

By doing this, data products that are created have high data trust from their initiation. When implemented correctly, they should be built with the enforcement of governance policies around access control, data definitions, system performance, and resource consumption. The Modern Data Stack by necessity enforces policies at query time, as data is accessed by different users. Proper access control services ensure data consumers get access to all the data they are entitled to see.

Business Goals for a Modern Data Stack

As a goal, a Modern Data Stack should enable greater data accessibility and business agility by addressing the entire data value chain from source to output. At its core, modernizing the data stack should be about industrializing data.

CIOs and CDOs who have earmarked investments for a Modern Data Stack say they want to turn data into a strategic asset. This means creating the ability to achieve one or more of the following business outcomes:

- Digital transformation and AI
- Enabling a data-driven business
- Gaining insights faster
- Unlocking the value of digital assets for innovation

Realizing these goals means the Modern Data Stack should address the mess, tackling tech debt and data silos, and enabling a data culture. The biggest challenge for legacy organizations has been “silos and spaghetti”—a tangled enterprise organizational structure with multiple departments performing similar functions, but failing to work together or share data. For this reason, Constellation Research’s VP and Principal Analyst Dion Hinchcliffe says, “Data leaders need a way to systematically create and manage a data fabric across all clouds, with local variation only occurring when required.”

Biggest challenge for legacy organizations?

“Silos and spaghetti.”

“Data leaders need a way to systematically create and manage a data fabric across all clouds, with local variation only occurring when required.”

DION HINCHCLIFFE
VP and Principal Analyst, Constellation Research

Broadening Considerations of the Modern Data Stack

We applaud that the Modern Data Stack considers the importance of data discovery, data governance, data observability, entitlements, and security. However, the challenge is these functions are neither separate nor capable of being provided by a single company or product.

Let's start with data discovery. Essentially, there are three types of data discovery that organizations need to accomplish:

1. Discovery of what data is out there
2. Discovery of what data should be classified as sensitive
3. Discovery of the origin and condition of the data

All three functions are equally important. The first makes data accessible for self-service discovery. The second facilitates the understanding of the presence and location of sensitive data needing protection. The third helps data users, data stewards, and data engineers understand the condition of data, where it came from, and how it has been processed or manufactured.

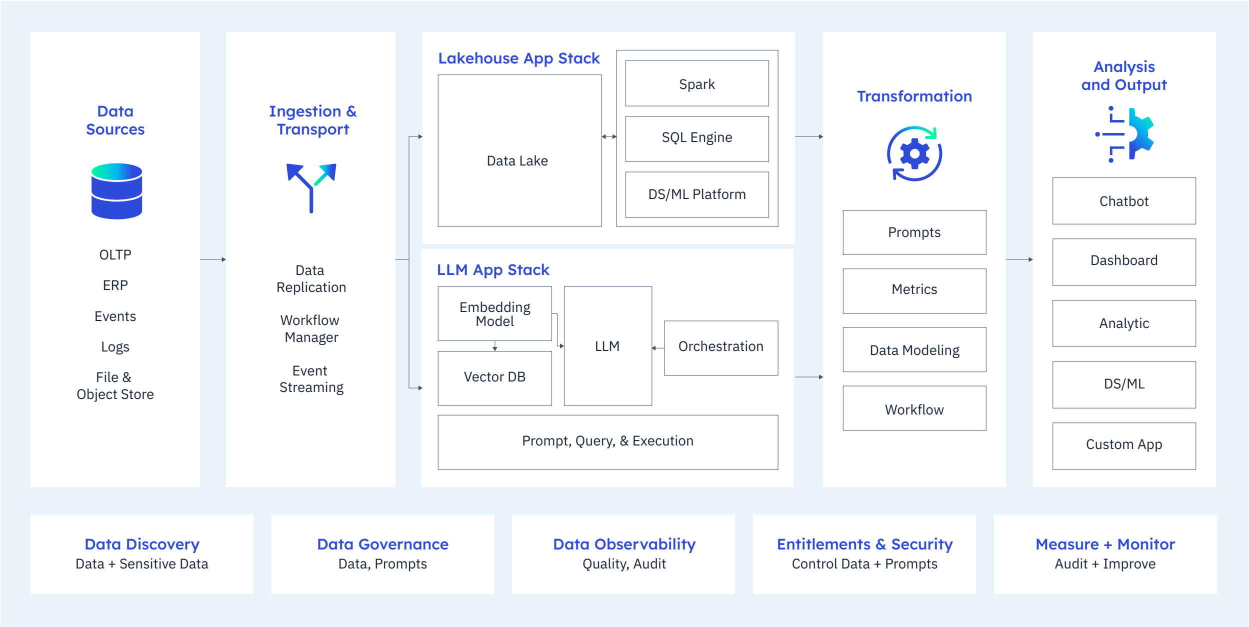
Additionally, data governance is not a standalone strategy. Data governance needs to be practiced for discovery, observability, entitlements, and security. For example, data catalogs and data observability systems should be governed in terms of how they create glossaries and access data, in addition to how they manage data quality. For this reason, policies are required for making data usable as well as for quality and testing. At the same time, entitlements at their core are about policy and controls. And while entitlement and control policies can originate in separate data governance solutions, why should policies be separated from their controls? This is critical in particular for compliance auditing.

At the same time, there needs to be a connection between policies, entitlements, security, and the ethical use of AI. For instance, AI model bias is dangerous. For this reason, organizations need to govern the data going into their AI models.

Finally, we call attention to a missing component we refer to as "manage and measure." Data leaders need this to understand how they are managing their data processes. As a goal, this should be improving over time. For this reason, we suggest a data-process-analysis block be added at the bottom of the above "Emerging Architectures for Modern Data Infrastructure" diagram. This should embody modern notions like DevOps, DataOps, and DataGovOps.

Adding Generative AI to the Mix

As was suggested earlier, the Modern Data Stack did not consider Generative AI. For this reason, the diagram by Bornstein, Li, and Casado needs to be extended. It needs to add the notion of an Large Language Model (LLM) App Stack as well as the notion of prompts and a Chatbot representing a self-service business intelligence (BI) output.



Modern data stacks must be more comprehensive, incorporating LLMs and enhancing interoperable data functions.

The updated diagram persists the notions of data discovery, data governance, data observability, entitlements, and security, adding a measure and management layer. Here, however, these notions need to be broadened. Data discovery needs to include not just data but sensitive data. Sensitive data needs to be discovered going into a LLM and being extracted in the form of prompts and prompt responses.

This means data governance needs to exist for data and prompts. Meanwhile, quality data is needed regardless of whether data is going into a data lake or vector database. Finally, entitlements and security need to be managed regardless of where data or data relationships are housed.

How Privacera Helps

Privacera helps organizations speed self-service BI and ensure data is secured regardless of whether it's being deployed in a cloud data lake, warehouse, or Generative AI LLM. Regardless of modality, Privacera eliminates IT as a Data Access Bottleneck and enables businesses to get appropriate data faster. Privacera does this by automating sensitive data discovery, creating global controls, and granting access by role and attribute. We can automate how data consumers receive data authorization from data queries, catalogs, or LLMs.

Privacera also speeds up the process for turning data policies into controls. Privacera does this by eliminating silos between data governance and controls. It also does this by integrating data governance and catalog tools with entitlements and security while improving traceability and auditability. At the same time, Privacera automates the process of creating controls with a no-code interface that goes from discovery to control and does so globally.

At the same time, Privacera empowers business applications deploying large language models. Today, most organizations are concerned with the risk of data leakage. For some, this means limiting or blocking access to LLMs. For others, they do not load sensitive data into their models. But that ultimately might result in less accurate, less useful models since they do not train the model on all available data, restricting authorized users from access to sensitive information they are, in reality, authorized to access. Essentially, Privacera allows organizations to add sensitive data to models and applications while providing fine-grained security and privacy controls to allow the right level of data disclosure based on user profiles and defined rules.

Privacera provides a single platform for controls and entitlements for traditional data systems and LLMs. The solution ensures only authorized data flows into LLMs and prompts release only appropriate data for each data consumer. Privacera accomplishes this by securing model inputs and outputs. Here, Privacera protects data exposure using context-aware data protection, inspecting user-prompted queries, and masking or redacting data requests containing improper or sensitive data. Meanwhile, Attribute-Based Access Control (ABAC) or Tag-Based Access Controls (TBAC) can be applied to mask sensitive data model output, ensuring users can only see data they are authorized to see.

This way, user prompts into the application and model are also inspected. Unauthorized questions that could expose sensitive data or are deemed toxic can be denied. Not only does this capability add an additional layer of security, but it also eliminates the massively expensive and wasteful LLM compute costs associated with processing unauthorized requests.

Lastly, Privacera improves your organization's security posture. Privacera does this with comprehensive compliance monitoring. This is provided with comprehensive dashboards and audit logs of all application and model activity, detailing what sensitive data is leveraged in each model, how it is protected, and who is accessing it.

Audit logs show:

- Who is accessing what models
- What sensitive data they are accessing
- When they accessed the model
- What protections were applied

Meanwhile, a security and compliance dashboard provides a view of the entire model landscape, including an overview of approved requests, denied requests, and requests that require masking to be applied. The dashboard also provides an overview of all sensitive data across models. These steps simplify model monitoring and compliance.

Conclusion

The modern stack represents a great reference architecture for businesses to bring data to the forefront of their organizations. To work for today's organization, it needs to add generative AI to the mix. As well, it needs to consider all aspects of the data value stream. Doing this creates a valuable model and approach for data leaders.

Fortune 500 enterprises trust Privacera for their universal data security, access control, and governance. Discover how to streamline data security governance with Privacera.

Take a unified approach to data access, privacy, and security with Privacera.

REQUEST A DEMO →   CONTACT US →

privacera