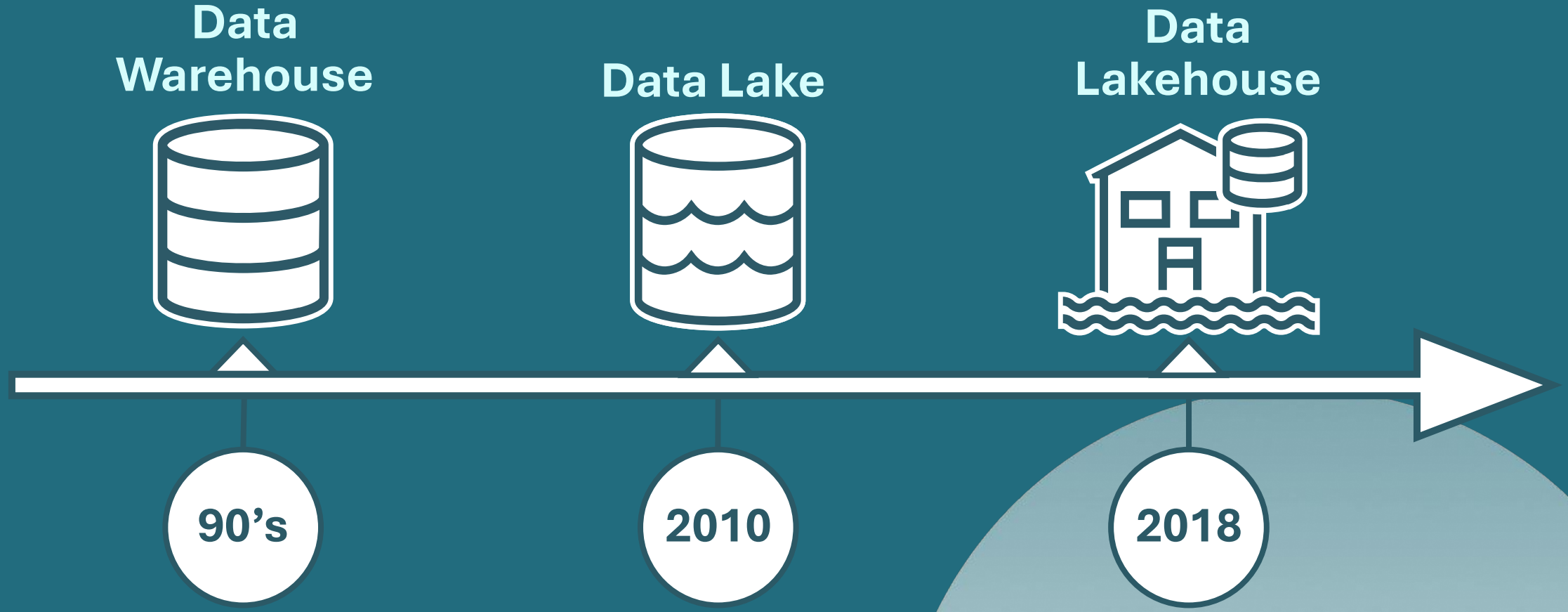
CONNECTED FUTURE 2024

Data Platform – Data Lakehouse in Oracle Cloud Infrastructure

Geert Zegers

Tripwire
solutions

Data Platform Emergence for Analytics



CONNECTED FUTURE
powered by Colabo

Tripwire solutions

Data Lakehouse

Data Warehouse + Data Lake = Data Lakehouse

Data Warehouse (the “house” in Lakehouse)

Emerged as a technology that brings together an organization’s collection of relational databases under a single umbrella, allowing the data to be queried and viewed as a whole.

- Stores processed and structured data, curated for a specific purpose, and stored in a specified format.
- Data typically queried by business users, used in analytics tools for reporting and projections.
- Typically includes data management features such as data cleansing and extract/load/transform (ETL).

Tripwire solutions

Why Not Run Analytics Against Your OLTP Environment?

Data warehouses are relational environments that are used for data analysis, particularly of historical data. Organizations use data warehouses to discover patterns and relationships in their data that develop over time.

- Their primary advantages are:
  - Integration of many data sources
  - Data optimized for read access
  - Ability to run quick ad hoc analytical queries
  - Data audit, governance and lineage
- In contrast, transactional environments are used to process transactions on an ongoing basis and are commonly used for order entry and financial and retail transactions. They do not build on historical data; in fact, in OLTP environments, historical data is often archived or simply deleted to improve performance.

Tripwire
solutions

Data warehouses and OLTP systems differ significantly

	Data Warehouse	OLTP
System Workload	Accommodates ad hoc queries and data analysis	Supports only predefined operations
Data modifications	Automatically updates on a regular basis	Updates by end users issuing individual statements
Schema design	Uses partially denormalized schemas to optimize performance	Uses fully normalized schemas to guarantee data consistency
Data scanning	Encompasses thousands to millions of rows	Accesses only a handful of records at a time
Historical data	Stores many months or years of data	Stores data for only weeks or months

Tripwire
solutions

What is a Cloud Data Warehouse?

A cloud data warehouse uses the cloud to ingest and store data from disparate data sources.

- The original data warehouses were built with on-premises appliance-based hardware. These on-premises data warehouses continue to have many advantages today. In many cases, they can offer improved governance, security, data sovereignty, and better latency.
- However, on-premises data warehouses are not as elastic and they require complex forecasting to determine how to scale the data warehouse for future needs. Managing these data warehouses can also be very complex.

Tripwire
solutions

Cloud Data Warehouse (CDW)

- Some of the advantages of cloud data warehouses include:
  - Elastic, scale-out support for large or variable compute or storage requirements
  - Ease of use
  - Ease of management
  - Cost savings
- The best cloud data warehouses are fully managed and self-driving, ensuring that even beginners can create and use a data warehouse with only a few clicks.
- An easy way to start your migration to a cloud data warehouse is to run your cloud data warehouse on-premises, behind your data center firewall which complies with data sovereignty and security requirements.
- In addition, most cloud data warehouses follow a pay-as-you-go model, which brings added cost savings to customers.

Tripwire
solutions

Data Sources

Oracle Applications

- ERP
- CRM
- Any Applications
- Social Media
- Digital Assets
- IoT

Capture data in structured or unstructured format

Oracle Autonomous Database

Processing	Insights	Development
Structured and Unstructured Data	Analysis / Visualization, Machine Learning	APEX, API

Transactions | Data Warehouse | Data Lake

1. Oracle Autonomous Database provides an easy-to-use, fully autonomous database that scales elastically and delivers fast query performance.
2. Autonomous Database breaks the barrier between transactional and analytical database and includes all your preferred services under one roof
3. Autonomous Database supports integrating with data lakes not just on Oracle Cloud Infrastructure, but also on Amazon, Azure, Google and more.

Outcomes

- Improved Real-Time Visibility
- Meaningful Business Insight
- Anomaly Detection Analysis
- Accurate Forecasts and Predictions

Oracle Machine Learning Notebooks are included in the platform.
A tight integration is completed with Oracle Analytics Cloud ensuring you get the outcomes you need right from a single fully managed platform

Data Lake
(the “lake” in Lakehouse)

About a decade ago companies began building data lakes

- Low-cost storage repository for structured, semistructured, and unstructured data in any format and size and at any scale that can be analyzed easily. Primarily used by data scientists, but also by business analysts, product managers, and other types of end users.
- It is a big data concept. Unstructured raw data from various organizational sources goes into the lake, often for staging prior to loading into a data warehouse and building data sets.
- Data lakes store an abundance of disparate, unfiltered data to be used later for a particular purpose. Data from line-of-business applications, mobile apps, social media, IoT devices, and more is captured as raw data in a data lake.

Tripwire
solutions

Data Lake
(the “lake” in Lakehouse)

While suitable for storing data, data lakes lack some critical features:

- Do not support transactions, they do not enforce data quality, and their lack of consistency / isolation makes it almost impossible to mix appends and reads, and batch and streaming jobs.
- For these reasons, many of the promises of the data lakes have not materialized, and in many cases leading to a loss of many of the benefits of data warehouses.
- The structure, integrity, selection, and format of the various datasets is derived at the time of analysis by the person doing the analysis.
- When organizations need low-cost storage for unformatted, unstructured data from multiple sources that they intend to use for some purpose in the future, a data lake might be the right choice.

Tripwire
solutions

Do I Need a Data Lake?

- Organizations use both data lakes and data warehouses for large volumes of data from various sources. The choice of when to use one or the other depends on what the organization intends to do with the data.
- Data warehouses are specifically intended to analyze data. Analytical processing within a data warehouse is performed on data that has been readied for analysis—gathered, contextualized, and transformed—with the purpose of generating analysis-based insights. Data warehouses are also adept at handling large quantities of data from various sources.
- When organizations need advanced data analytics or analysis that draws on historical data from multiple sources across their enterprise, a data warehouse is likely the right choice.

Tripwire
solutions

The Best of Both Worlds

Data Warehouse

- Relational Storage
- Structured data
- ACID
- Primarily SQL

Data Lake

- Stores files
- (Non-)structured data
- No ACID
- SQL, Python, Scala, ...

Data Lakehouse

- Stores files
- (Non-)structured data
- ACID
- SQL, Python, Scala, ...

Tripwire
solutions

Oracle Autonomous Data Warehouse (ADW) and Data Lakehouse Integration

Offers several advantages for organizations seeking to maximize the potential of their diverse data sources.

- Integration combines the strengths of a traditional data warehouse with the scalability and flexibility of a data lake, providing a unified platform for storing, processing and analyzing structured and unstructured data.
- Enables businesses to break down data silos, providing a holistic view of information and facilitating more comprehensive analytics.
- Additionally, organizations can leverage the power of SQL-based analytics for structured data stored in the data warehouse while accommodating the storage and processing of vast amounts of semi-structured and unstructured data in the data lake.
- This approach supports a wider range of analytics use cases, from traditional business intelligence reporting to advanced analytics and machine learning.

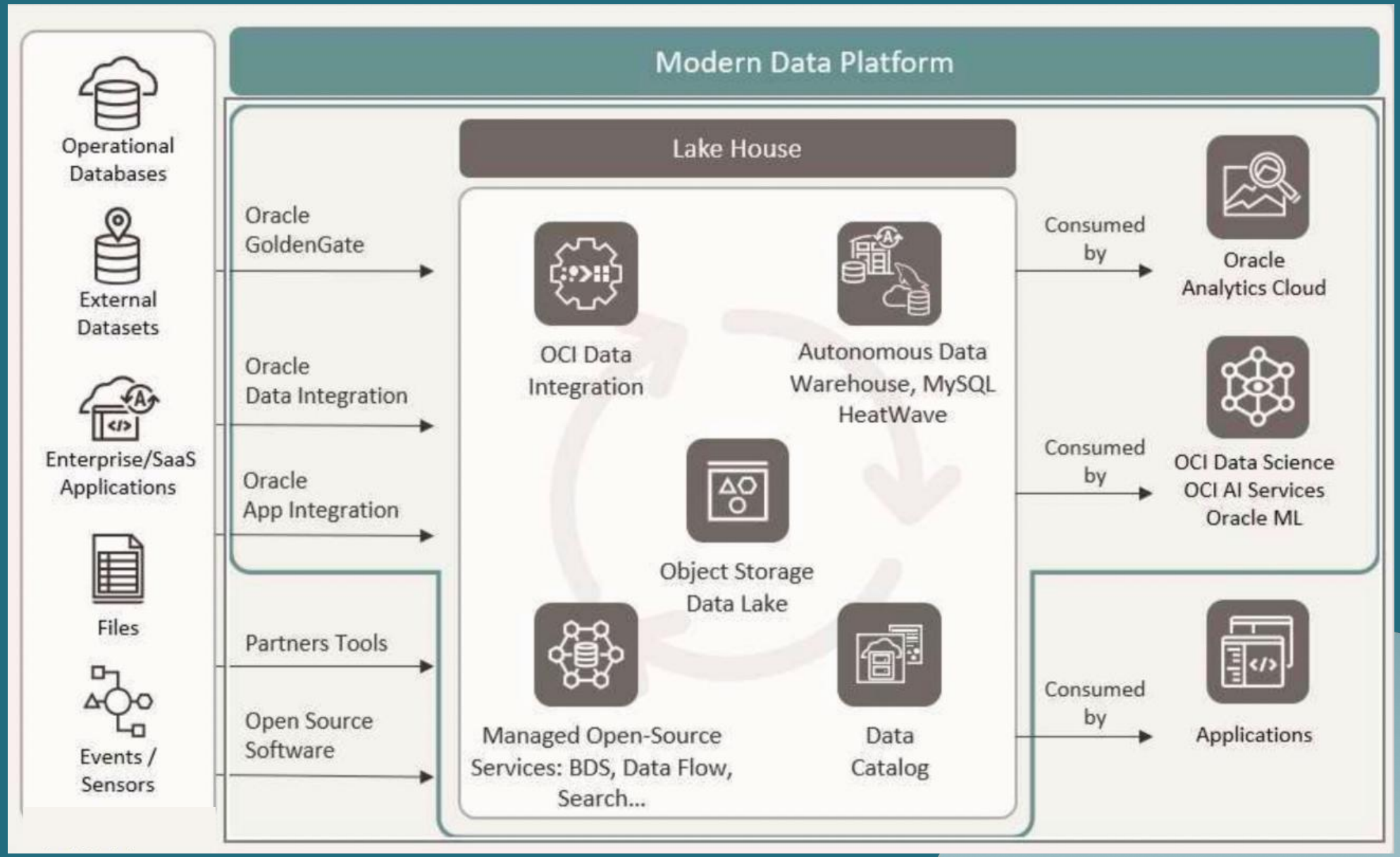
Tripwire
solutions

Oracle Cloud Infrastructure (OCI) Services

- With Oracle Cloud Infrastructure (OCI), you can build a secure, cost-effective, and easy-to-manage data lake. A data lake on OCI is tightly integrated with your preferred data warehouses and analytics as well as with other OCI services.

- Improves data agility, allowing businesses to adapt quickly to changing requirements, explore new data sources, and derive deeper insights from their data ecosystem.

Tripwire
solutions

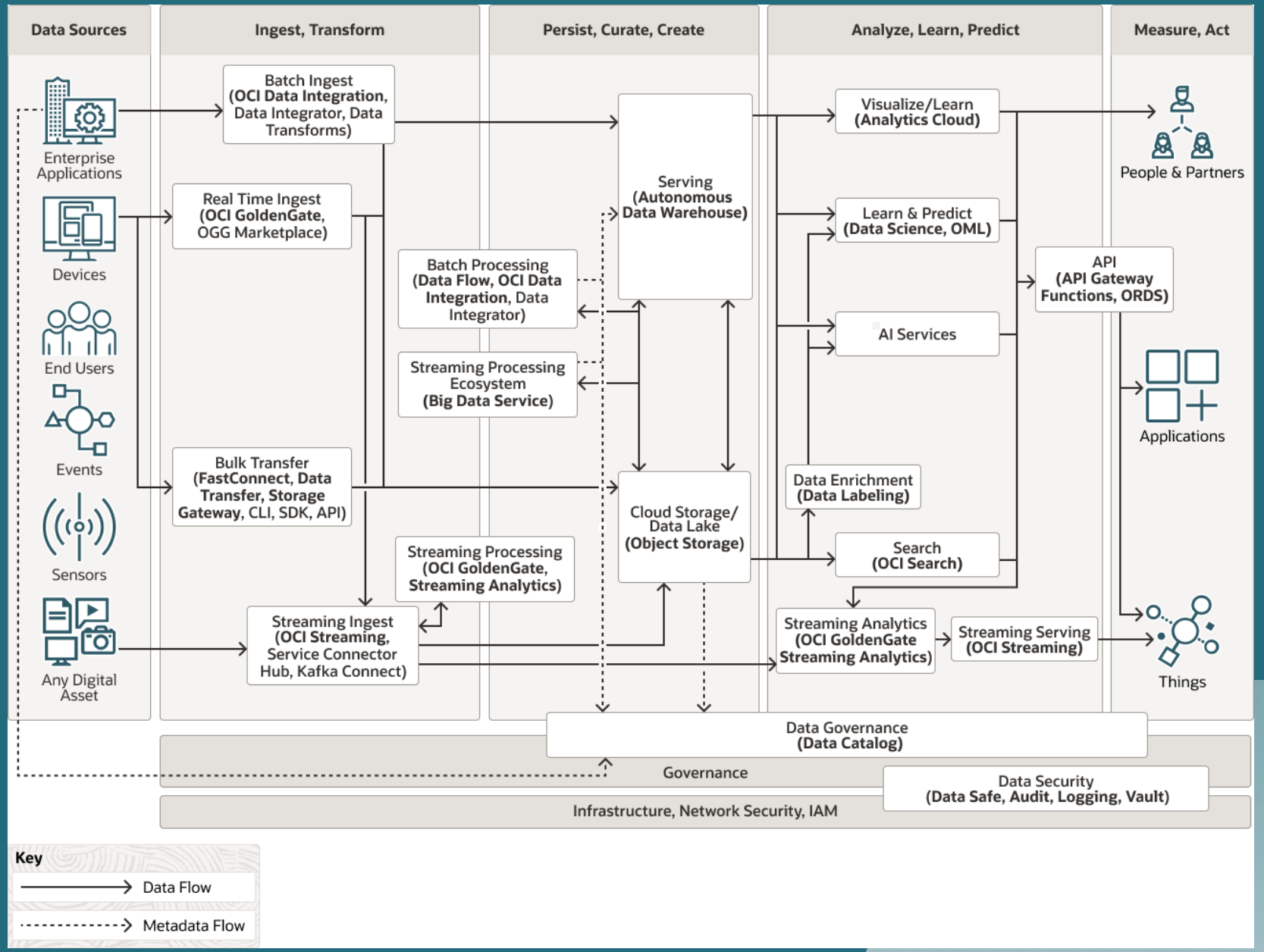

Modern Data Platform
Operational Databases
External Datasets
Enterprise/SaaS Applications
Files
Events / Sensors
Lake House
Oracle GoldenGate
Oracle Data Integration
Oracle App Integration
Partners Tools
Open Source Software
OCI Data Integration
Autonomous Data Warehouse, MySQL HeatWave
Object Storage Data Lake
Managed Open-Source Services: BDS, Data Flow, Search...
Data Catalog
Consumed by
Oracle Analytics Cloud
Consumed by
OCI Data Science
OCI AI Services
Oracle ML
Consumed by
Applications

Best practices for integrating Oracle ADW with your Data Lakehouse

Understand Data Lakehouse Architecture

- To successfully integrate Oracle ADW with your Data Lakehouse, it's crucial to understand the underlying architecture.
- A data lakehouse combines the strengths of a data warehouse and a data lake, allowing for unified storage and processing of structured and semi-structured data.
- Understanding this architecture aids in designing effective data pipelines, ensuring optimal data storage, and enabling efficient query processing across the integrated platforms.
- However, the integration process involves connecting and managing data across different storage and processing platforms.

Tripwire
solutions


Data Sources
Ingest, Transform
Persist, Curate, Create
Analyze, Learn, Predict
Measure, Act
Enterprise Applications
Devices
End Users
Events
Sensors
Any Digital Asset
Batch Ingest (OCI Data Integration, Data Integrator, Data Transforms)
Real Time Ingest (OCI GoldenGate, OGG Marketplace)
Bulk Transfer (FastConnect, Data Transfer, Storage Gateway, CLI, SDK, API)
Streaming Ingest (OCI Streaming, Service Connector Hub, Kafka Connect)
Batch Processing (Data Flow, OCI Data Integration, Data Integrator)
Streaming Processing Ecosystem (Big Data Service)
Streaming Processing (OCI GoldenGate, Streaming Analytics)
Serving (Autonomous Data Warehouse)
Cloud Storage/ Data Lake (Object Storage)
Visualize/Learn (Analytics Cloud)
Learn & Predict (Data Science, OML)
AI Services
Data Enrichment (Data Labeling)
Search (OCI Search)
Streaming Analytics (OCI GoldenGate Streaming Analytics)
Streaming Serving (OCI Streaming)
API (API Gateway Functions, ORDS)
People & Partners
Applications
Things
Data Governance (Data Catalog)
Governance
Data Security (Data Safe, Audit, Logging, Vault)
Infrastructure, Network Security, IAM
Key
Data Flow
Metadata Flow

Use Compatible Data Formats

Appropriate data formats are critical to seamless integration between Oracle ADW and your data lake.

- Advisable to choose formats like Parquet or ORC for structured data and Avro or JSON for semi-structured or unstructured data.
- These widely supported formats provide compatibility across different storage systems and facilitate efficient data processing and querying.
- Ensuring compatibility enhances interoperability and allows for smooth data exchange between the data lakehouse components.

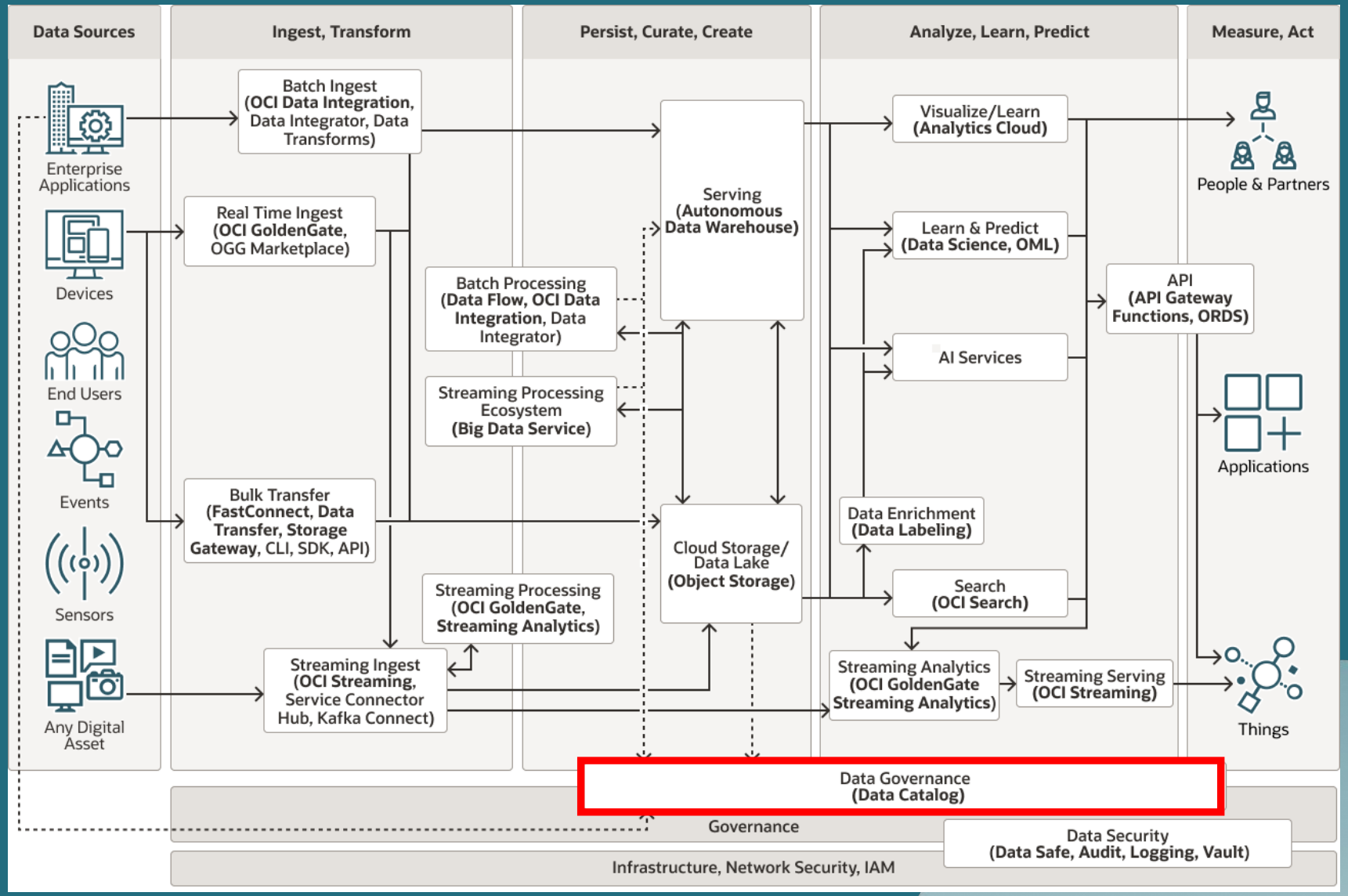
Tripwire
solutions

Optimize Data Lake Storage

Optimizing storage in your data lake is essential for performance and cost considerations.

- Leverage features such as partitioning, compression, and clustering to organize and store data efficiently.
- Partitioning allows logical data segregation, facilitating faster query performance, while compression reduces storage space requirements. Clustering organizes data physically, enhancing retrieval speed.
- By employing these techniques, you can enhance the overall data storage efficiency in the data lakehouse, leading to improved performance and reduced storage costs.

Tripwire
solutions



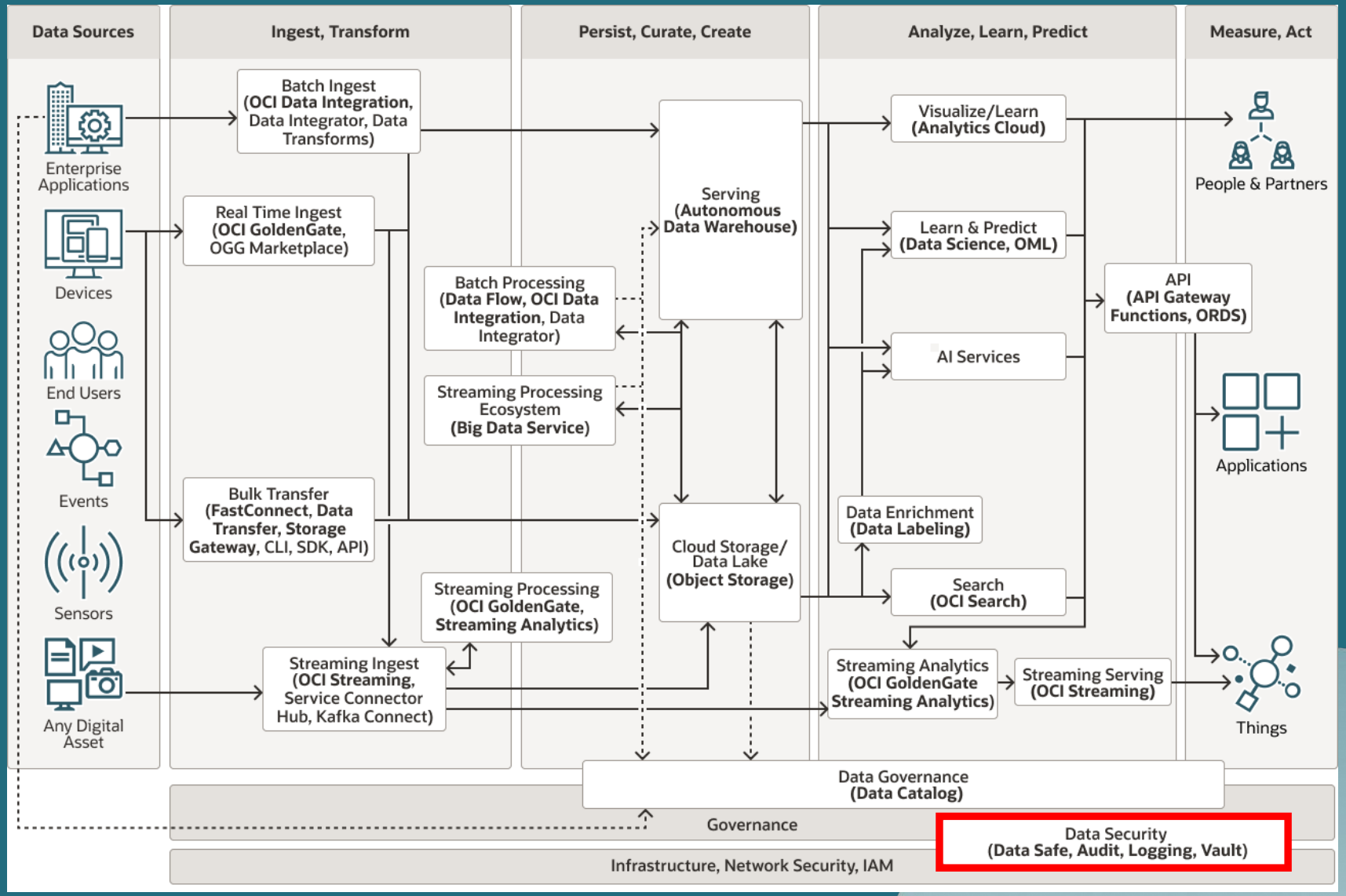
Tripwire solutions

Implement Data Catalogs

Successful integration of Oracle ADW with your data lakehouse enabled by the implementation of data catalogs.

- These catalogs act as centralized metadata repositories, providing comprehensive information about the data stored in both ADW and the data lake.
- Maintaining a detailed catalog allows users to discover, understand, and govern their data assets quickly.
- This includes information on data lineage, quality, and usage, contributing to improved data governance and fostering a more transparent and collaborative data environment.

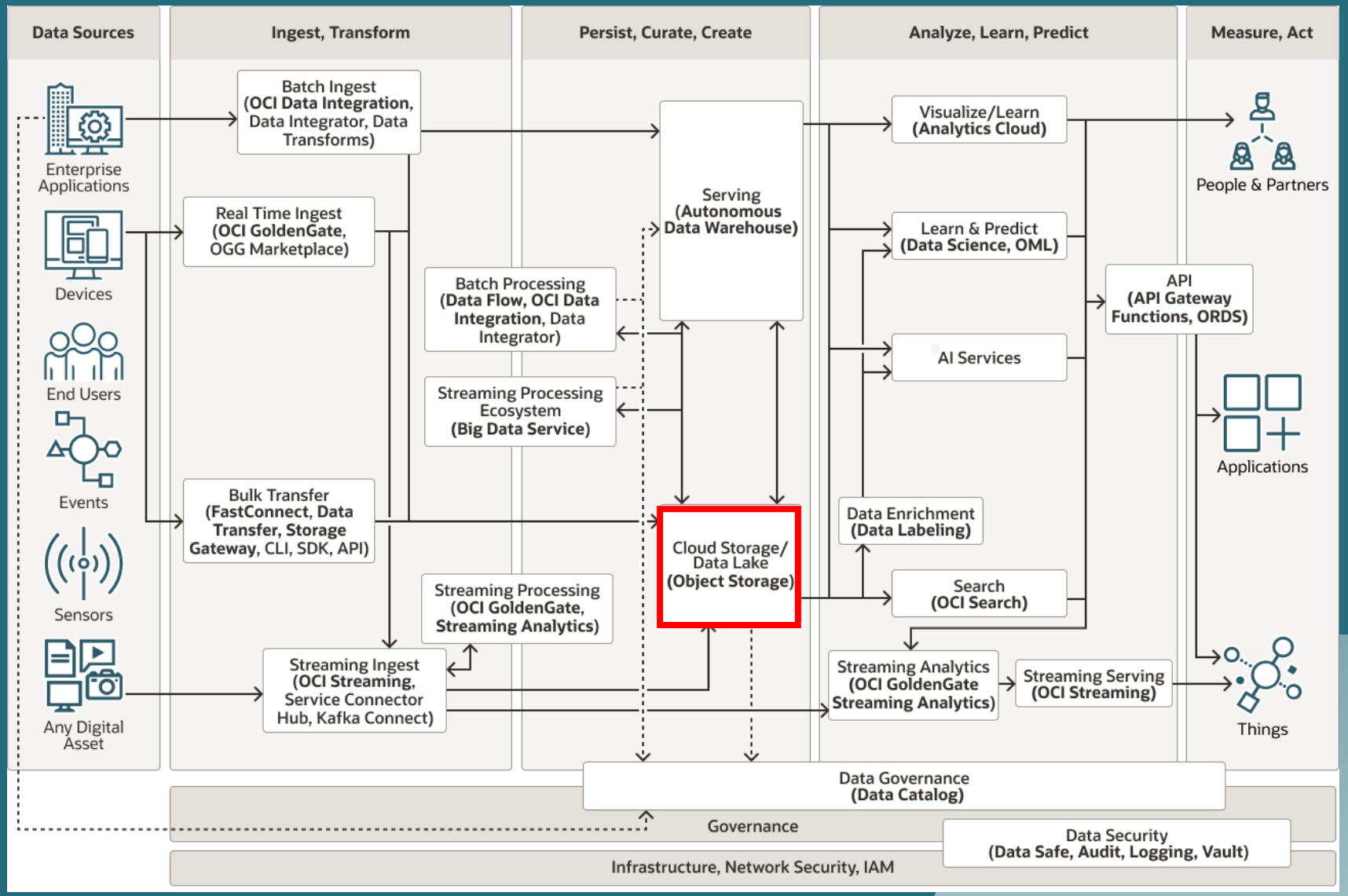
Tripwire
solutions


Data Sources
Ingest, Transform
Persist, Curate, Create
Analyze, Learn, Predict
Measure, Act
Enterprise Applications
Devices
End Users
Events
Sensors
Any Digital Asset
Batch Ingest (OCI Data Integration, Data Integrator, Data Transforms)
Real Time Ingest (OCI GoldenGate, OGG Marketplace)
Bulk Transfer (FastConnect, Data Transfer, Storage Gateway, CLI, SDK, API)
Streaming Ingest (OCI Streaming, Service Connector Hub, Kafka Connect)
Batch Processing (Data Flow, OCI Data Integration, Data Integrator)
Streaming Processing Ecosystem (Big Data Service)
Streaming Processing (OCI GoldenGate, Streaming Analytics)
Serving (Autonomous Data Warehouse)
Cloud Storage/ Data Lake (Object Storage)
Visualize/Learn (Analytics Cloud)
Learn & Predict (Data Science, OML)
AI Services
Data Enrichment (Data Labeling)
Search (OCI Search)
Streaming Analytics (OCI GoldenGate Streaming Analytics)
Streaming Serving (OCI Streaming)
API (API Gateway Functions, ORDS)
People & Partners
Applications
Things
Data Governance (Data Catalog)
Governance
Data Security (Data Safe, Audit, Logging, Vault)
Infrastructure, Network Security, IAM

Secure Data Access

- Utilize Oracle Cloud Infrastructure (OCI) Identity and Access Management (IAM) to implement robust access controls, ensuring that only authorized users and applications have the necessary permissions to access and modify data.
- By employing proper authentication and authorization mechanisms, organizations can safeguard sensitive information, mitigate the risk of unauthorized access, and maintain compliance with data privacy regulations.
- Implement security features like Encryption both at rest and in transit and Fine-grained data access control both at row and column level

Tripwire
solutions



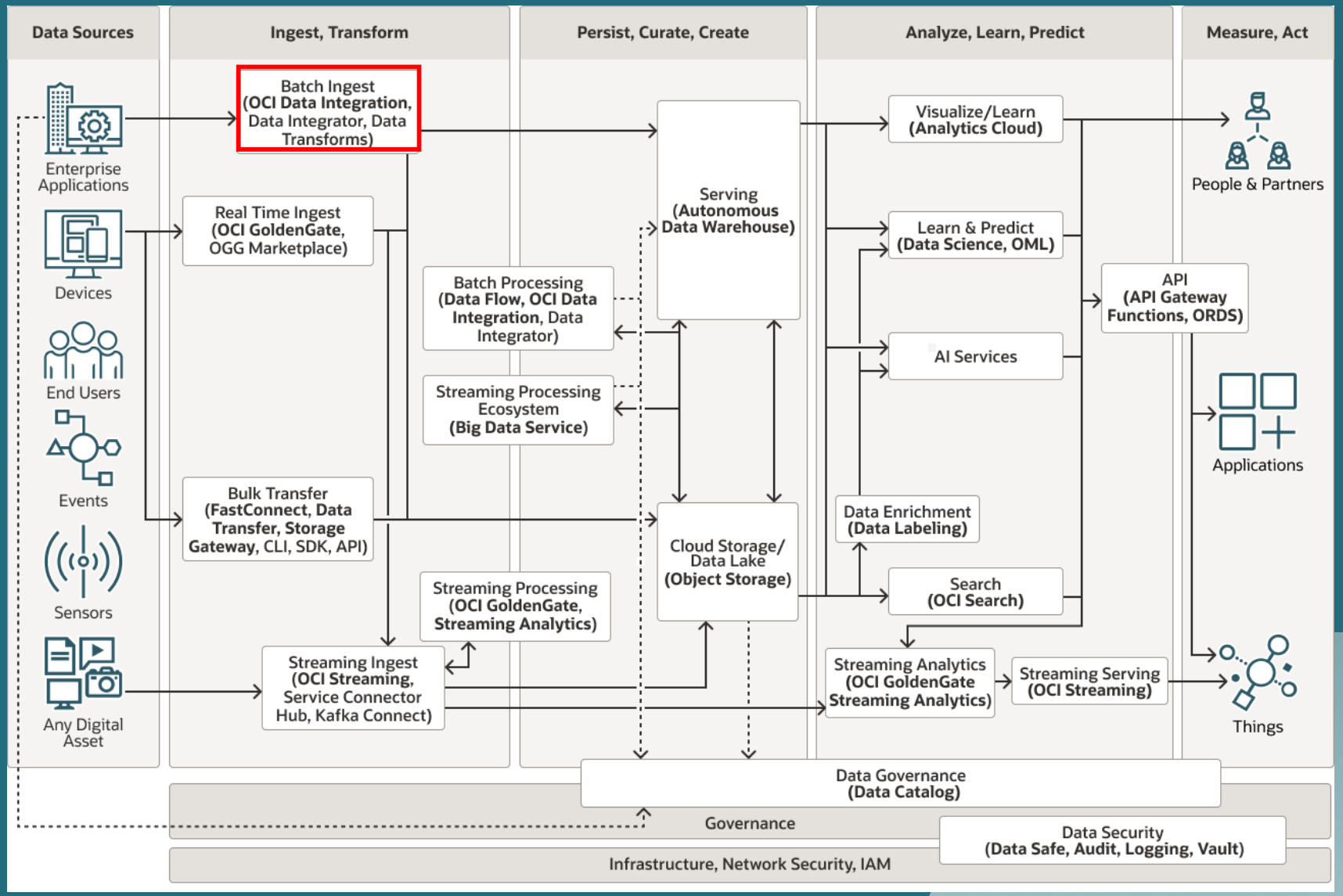
Tripwire solutions

Leverage Oracle Cloud Services

- Integrating Oracle ADW with your data lakehouse can be further optimized by leveraging additional Oracle Cloud services.

- Oracle Cloud Object Storage provides scalable and durable storage for large volumes of data.

Tripwire
solutions



Tripwire solutions

Use Data Integration Tools

Streamlining data movement between Oracle ADW and your data lakehouse is essential for efficient integration.

- Oracle provides powerful data integration tools to simplify the ETL processes.

  Oracle Data Integration services, such as Oracle Data Integrator (ODI) and Oracle Cloud Data Flow, streamline the ETL processes of moving data between platforms.
- These tools enable organizations to design, schedule, and manage complex data workflows, ensuring that data is extracted from source systems, transformed according to business requirements, and seamlessly loaded into the target systems.
- Leveraging these tools simplifies the integration process, enhances data quality, and contributes to the overall effectiveness of data processing workflows.
- By incorporating these complementary services, organizations can enhance scalability, performance, and overall efficiency in managing and processing their integrated data.

Tripwire
solutions

Ensure Data Consistency and Quality

Maintaining data consistency and quality is critical to integrating Oracle ADW with your data lakehouse.

- Implement robust data quality checks and validation processes to ensure the integrated data is accurate, complete, and meets the defined standards. Regular monitoring and cleansing activities should be performed to identify and rectify any discrepancies or anomalies.
- By prioritizing data quality, organizations can build trust in the integrated data, support reliable decision-making processes, and reduce the risk of errors that may arise from inconsistent or inaccurate information.

Tripwire
solutions

Optimize Query Performance

To enhance the overall efficiency of querying data across Oracle ADW and the data lake, it is essential to optimize SQL queries.

- Consider utilizing partition pruning, indexing, and optimizing join operations.

  Partition pruning involves minimizing the data scanned by the database engine by selecting only relevant partitions, thereby improving query response times.
- Proper indexing ensures faster data retrieval, and optimizing join operations contributes to efficient query processing.
- Organizations can improve responsiveness and user experience when querying integrated data across different storage platforms by incorporating these performance optimization techniques.

Tripwire
solutions

Monitor and Maintain

Establishing robust monitoring mechanisms is crucial for the ongoing health and performance of the integration between Oracle ADW and your data lakehouse.

- Implement monitoring solutions to track key metrics such as query performance, data processing times, and storage utilization. Set up alerts for potential issues or anomalies, allowing for proactive intervention.
- Regular maintenance tasks, including updating statistics, optimizing storage structures, and performing routine checks, should be carried out to ensure the continued efficiency and reliability of the integrated data environment. This proactive approach enables organizations to identify and address issues promptly, minimizing potential disruptions to data workflows.

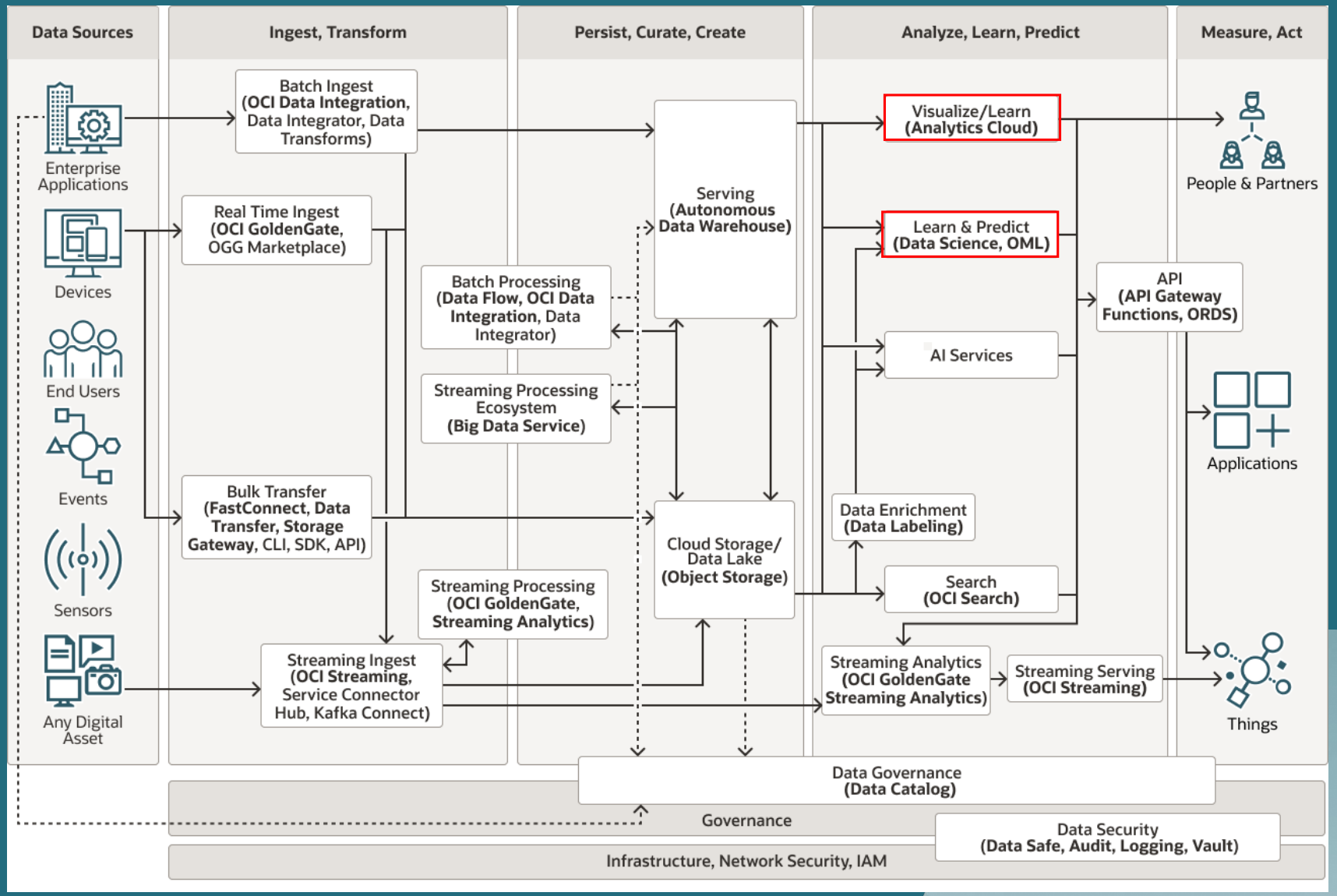
Tripwire
solutions

Document Integration Processes

Comprehensive documentation of the integration processes is fundamental for the long-term success and sustainability of the Oracle ADW and data lakehouse integration.

- Document data flows, transformations, dependencies, and any custom scripts or configurations used in the integration.
- This documentation is valuable for troubleshooting, onboarding new team members, and ensuring consistency in data management practices. Additionally, it provides insights into the rationale behind specific design choices, aiding in future modifications or upgrades to the integration setup.
- Regularly updating and maintaining this documentation ensures the integration remains well-documented, transparent, and easily manageable throughout its lifecycle.

Tripwire
solutions


Data Sources
Ingest, Transform
Persist, Curate, Create
Analyze, Learn, Predict
Measure, Act
Enterprise Applications
Devices
End Users
Events
Sensors
Any Digital Asset
Batch Ingest (OCI Data Integration, Data Integrator, Data Transforms)
Real Time Ingest (OCI GoldenGate, OGG Marketplace)
Bulk Transfer (FastConnect, Data Transfer, Storage Gateway, CLI, SDK, API)
Streaming Ingest (OCI Streaming, Service Connector Hub, Kafka Connect)
Batch Processing (Data Flow, OCI Data Integration, Data Integrator)
Streaming Processing Ecosystem (Big Data Service)
Streaming Processing (OCI GoldenGate, Streaming Analytics)
Serving (Autonomous Data Warehouse)
Cloud Storage/ Data Lake (Object Storage)
Visualize/Learn (Analytics Cloud)
Learn & Predict (Data Science, OML)
AI Services
Data Enrichment (Data Labeling)
Search (OCI Search)
Streaming Analytics (OCI GoldenGate Streaming Analytics)
Streaming Serving (OCI Streaming)
API (API Gateway Functions, ORDS)
People & Partners
Applications
Things
Data Governance (Data Catalog)
Governance
Data Security (Data Safe, Audit, Logging, Vault)
Infrastructure, Network Security, IAM

Modern data platforms OCI components

- Data Warehouses: Platforms like Oracle Autonomous Data Warehouse provide scalable and high-performance data warehousing capabilities for structured data storage and analysis.
- Data Lakes: Technologies such as the Oracle OCI Lake; Oracle Autonomous Database/Warehouse, Object Storage, Big Data; Apache Hadoop, Apache Spark, Oracle Data Flow; Data Catalog; Query Service,etc enable the storage and processing of large volumes of raw, semi-structured, and unstructured data.
- ETL (Extract, Transform, Load) Tools: Platforms like Oracle Data Integrator facilitate data ingestion, transformation, and loading processes, ensuring data quality and integration across different sources.
- Analytics and Visualisation Tools: Solutions like Oracle Analytics Cloud, Fusion Analytics Warehouse (for Oracle Fusion Applications) Look to provide intuitive interfaces for exploring and visualising data, enabling users to gain insights and communicate findings effectively.

Tripwire
solutions

Conclusions

- Data Lakehouses are an evolution, not a revolution
- Architects & Marketing sell the prospect of
  - self-service for everyone
  - lower cost
- great business opportunities
- But integration effort and resource consumption are a point of attention
- In-database computing on the other side
  - saves resources
  - eases development
  - scales well
- Combining both worlds using OCI services is a straightforward and comprehensive approach

Tripwire
solutions

Q&A

Tripwire
solutions