MCKNIGHT
CONSULTING GROUP

Data Integration – Newsflash: We Still Just Move Data!

Presented by: William McKnight

"#1 Global Influencer in Big Data" Thinkers360

President, McKnight Consulting Group

3 X Inc 5000

/in/wmcknight

www.mcknightcg.com
(214) 514-1444

Informatica

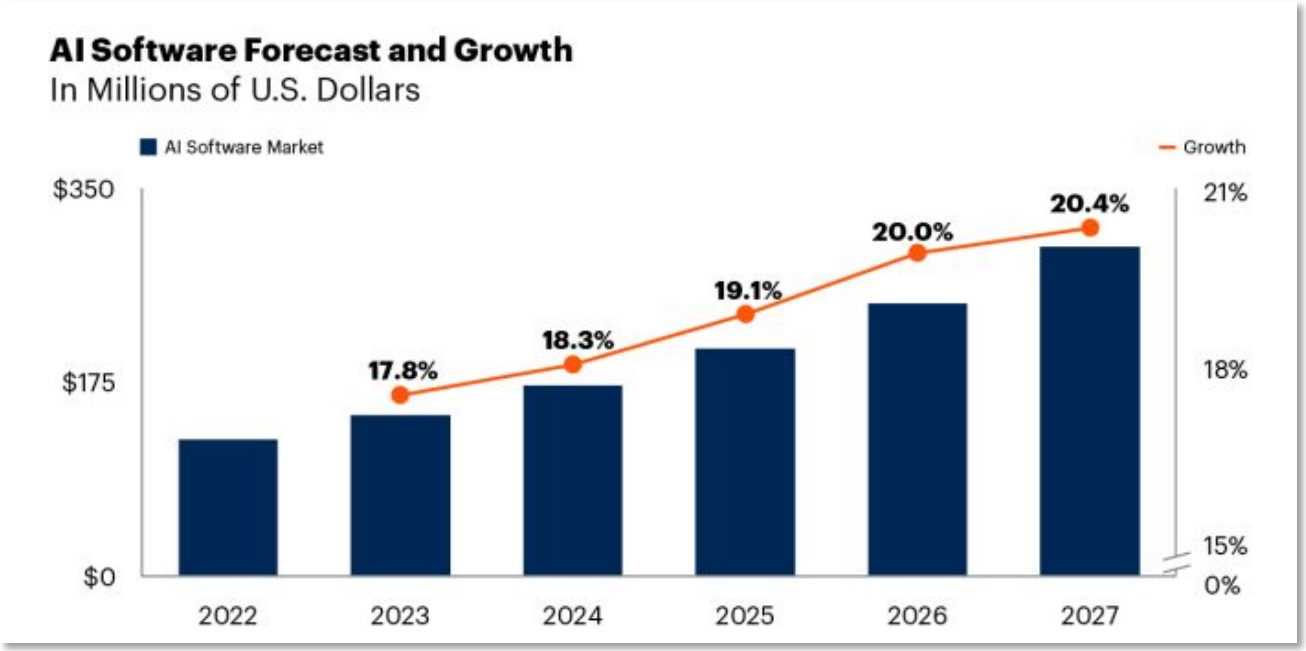
Data Engineering for AI

Preetam Kumar

Director, Product Marketing

Where data & AI come to LIFE

AI Addressable Market is Growing Exponentially



- By 2027, spending on AI software will grow to **$297.9 billion** with a CAGR of **19.1%.**
- Over the next five years, the market growth will accelerate from 17.8% to reach **20.4%** in 2027.
- Generative AI software spend will rise from 8% of AI software in 2023 to **35%** by 2027.

NOTE: **[1]** "Forecast Analysis: Artificial Intelligence Software,2023-2027, Worldwide." (2023), *Gartner.com*

The Emergence of GenAI has Changed the AI Strategy

	Current AI Strategy	Updated AI Strategy
Vision	AI automating tasks	Generative AI augmenting people in their work
Roadmap	3-year outlook, business innovation	1-year outlook, business criticality
Use Cases	Predictive analytics, automating tasks	Generating artifacts and simulate decisions (text, video, audio, code, data)
Governance	Fragmented	Clear business responsibility, AI ethics committee
Talent	AI center of excellence	Educate everybody on responsible use of GenAI

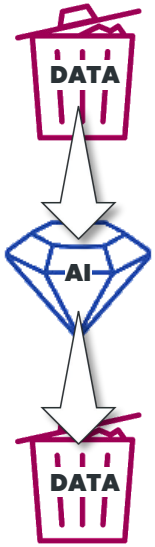
AI Needs Data Engineering and Data Engineering Needs AI

Success of AI models is dependent on the availability of trusted and timely data.

- Missing, incomplete, or inaccurate data adversely impact model's behavior
- Leads to incorrect or biased predictions and reduce the value and RoI

Data engineering needs AI

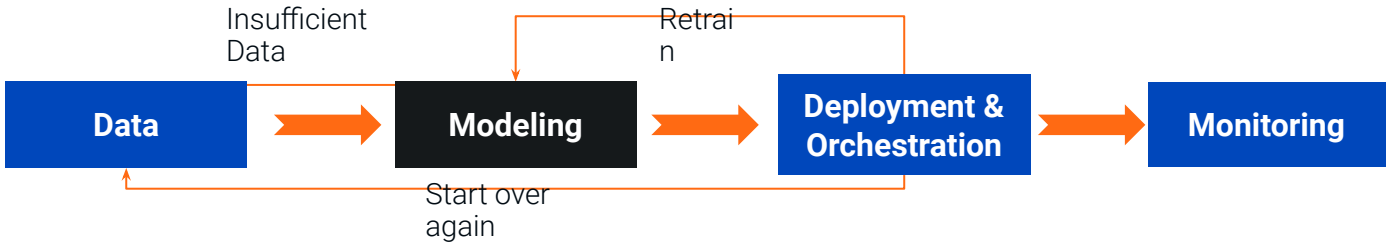
- Build data pipelines automatically based on user intent
- Autonomously optimize and control your cloud data engineering costs
- Autonomously identify & fix operational bottlenecks



© Informatica. Proprietary and Confidential.

Data Engineering for AI

Informatica is integral to the Data Architecture for AI



- Define data
- Procure data
- Data cleansing
- Data preparation
- Metadata management
- Data Governance

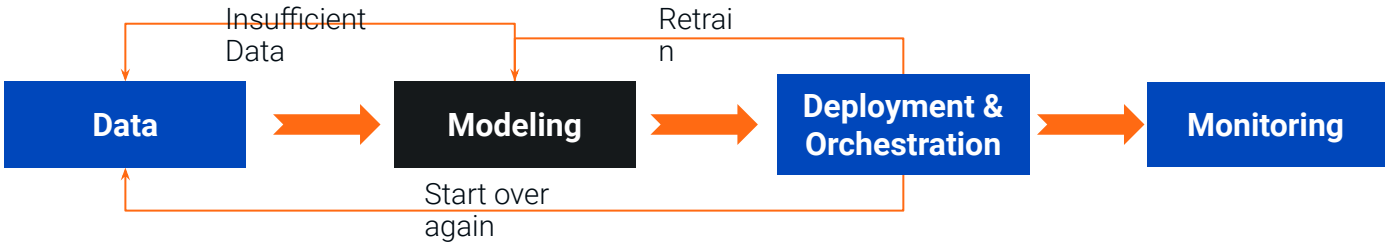
- Model selection
- Model training

- Model deployment
- Model consumption

- Model performance

Customer Stories for AI

Informatica is integral to the Data Architecture for AI

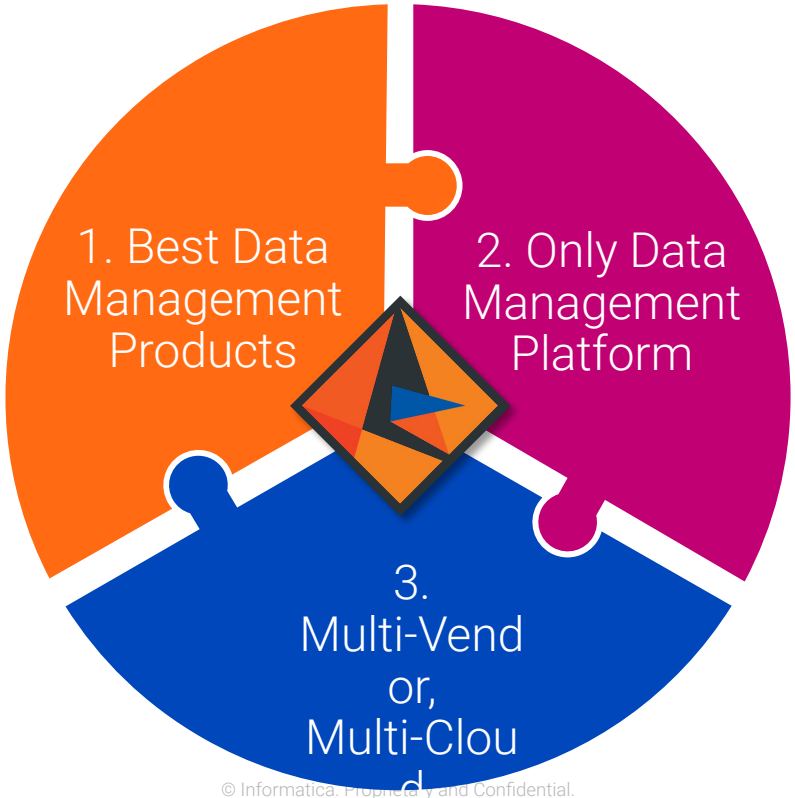


- SparkCognition Democratizes AI
  - "Informatica gives our customers many options for leveraging more of their enterprise data for AI, which will ultimately make their efforts more successful."
  - Simplified data integration, enabling users to incorporate more data sources and generate highly accurate and useful results from models built using Darwin™
  - Used Informatica Cloud Application Integration and Cloud Data Integration to capture streaming data for use in ML models
  - Allowed customers to integrate data from their source systems to feed machine learning (ML) models

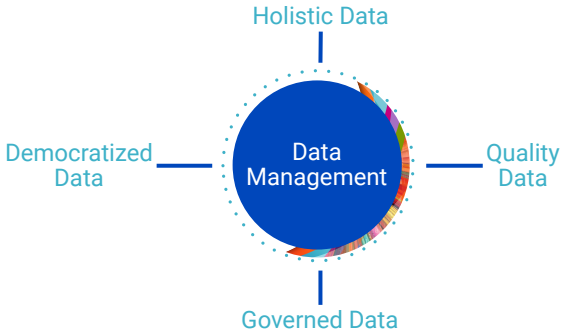
- Big pension firm at Canada using Informatica's iPaaS to orchestrate openAI models to build business processes
- Global Bank in APJ used Informatica's iPaaS to orchestrate Google BERT models to build FinServ business processes
- AI operationalization: Informatica IT moved from AzureML to Informatica's ModelServe and saved almost 13x cost

Informatica Differentiators

The Data Management Choice for Enterprise AI



Informatica for GenAI

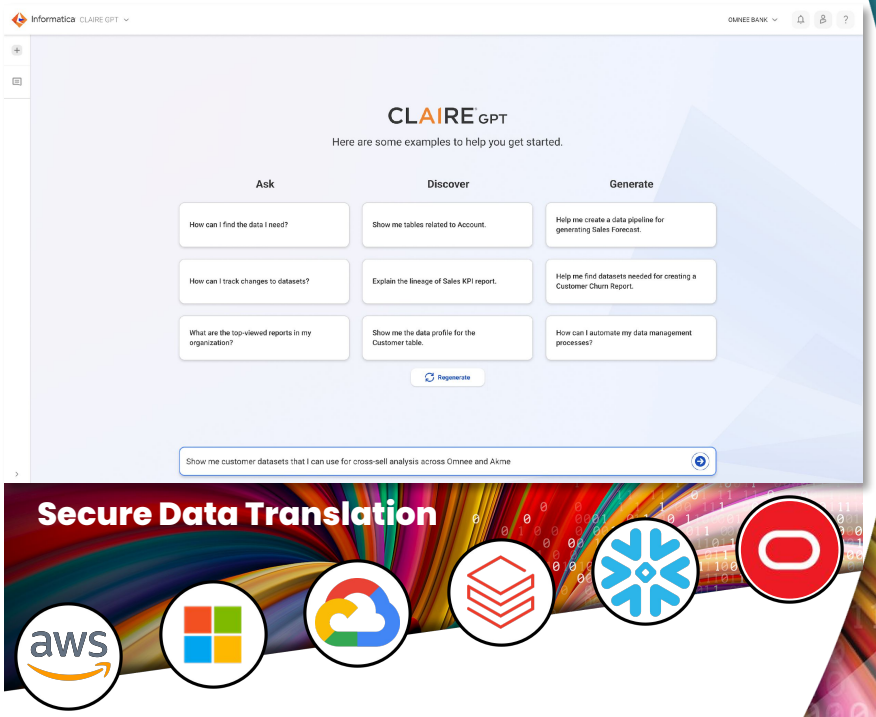


GenAI from Informatica

Sign Up for CLAIRE® GPT Free

Private Preview

AI Powered Insights



Across Multiple Data Sources

Informatica

vbelur@informatica.com

Thank You

Where data & AI come to LIFE

Where data & AI come to

Informatica

Partial Client List

CONSUMER PRODUCTS/RETAIL

FINANCIAL

INSURANCE/HEALTHCARE

PUBLISHING

OTHER

GOVERNMENT AND UTILITIES

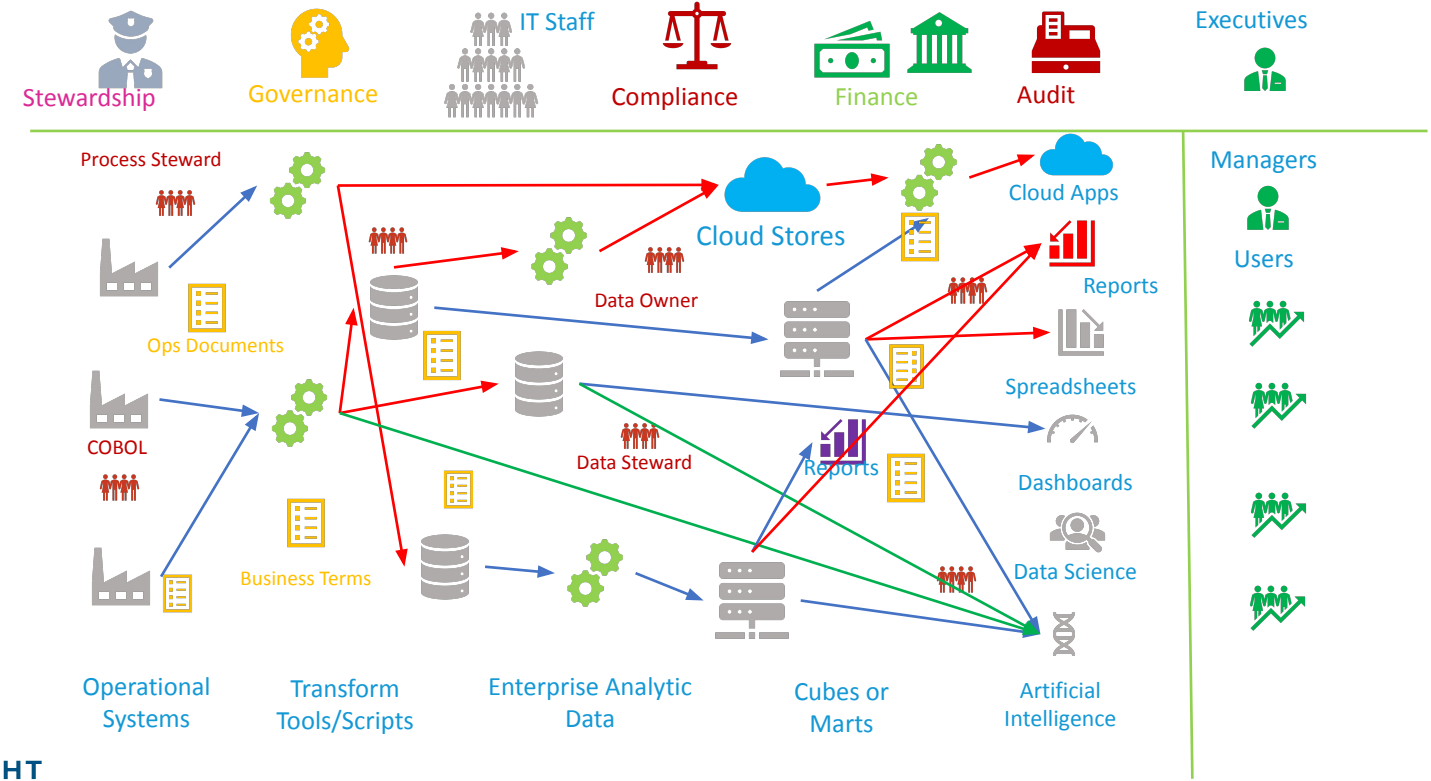
EDUCATION

PHARMACEUTICAL

TELECOMMUNICATIONS

Why So Many Data Stores?

Data Over Time



MCKNIGHT CONSULTING GROUP

Price-Performance

- Performance is a critical point of interest when it comes to selecting a platform
  - To measure data warehouse performance, we use similarly priced specifications across data warehouse competitors.
- Usually when people say they care about performance, it is the ultimate metric of price/performance
- In trying to understand Price-Performance, the realities of creating fair tests can be overwhelming to many shops, and is a task usually underestimated

Cost Predictability and Transparency

- The cost profile options for cloud databases are straightforward if you accept the defaults for simple workload or proof-of-concept (POC) environments.
- Initial entry costs and inadequately scoped environments can artificially lower expectations of the true costs of jumping into a cloud data warehouse environment.
- For some, you pay for compute resources as a function of time, but you also choose the hourly rate based on certain enterprise features you need.
- With some platforms, you pay for bytes processed and the underlying architecture is unknown. The environment is scaled automatically without affecting price. There is also a cost-per-hour flat rate where you would need to calculate how long it would take to run your queries to completion to predict costs.
- Customers need to analyze current workloads, performance, and concurrency and project those into realistic pricing in alternative platforms.

Administration

- Overall costs, time, as well as storage and compute resources are affected by the simplicity of configurability and overall use
- The platform should have embraced a self-sufficiency model for its customers and be well into the process of automating repetitive tasks
- Easy administration starts with setup that is a simple process of asking basic information and providing helpful information for selecting the storage and node configurations
- The data store should support mission-critical business applications with minimal downtime

Optimizer

- Conditional parallelism and what the causes are of variations in the parallelism deployed
- Dynamic and controllable prioritization of resources for queries
- Time requirements for optimal queries, such as compiling indexes or updating statistics
- Workload isolation capabilities

Concurrency Scaling

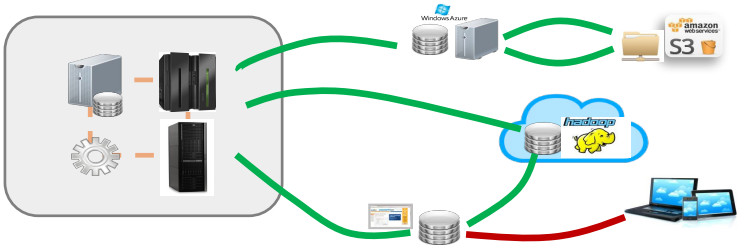
- If the database has concurrency limitations, designing around them is difficult at best, and limiting to effective data usage
- If the data store automatically scales up to overcome concurrency limitations, this may be costly if the data warehouse charges by compute node
- If the data store charges per user, costs will also increase as the data warehouse is put to more use in the company
- The workload may need linear scaling in overall query workload performance as concurrent users are added

Resource Elasticity

- You may need the ability to scale up and down and take advantage of the elastic compute and storage capabilities in the cloud, public or private, without disruption or delay
- The more the customer needs to be involved in resource determination and provisioning, the less elastic, and less modern, the solution is
- One thing to watch for in elasticity scaling is keeping the amount of money spent by the customer under the customer's control

Machine Learning

- Today, some data query languages need to be extended to include machine learning, or firms may find the programming required will be too challenging to keep pace
- Data stores today may need to weave machine learning into their data processing workflows
- Vendors must accommodate and extend SQL to include machine learning functions and algorithms to expand the capabilities of those tools and users
- If your database does not include machine learning, there are many extra things to be concerned with
- Other components will be needed to complete the toolbox and get the job done
- Ideally, security for machine learning will be the same as database security

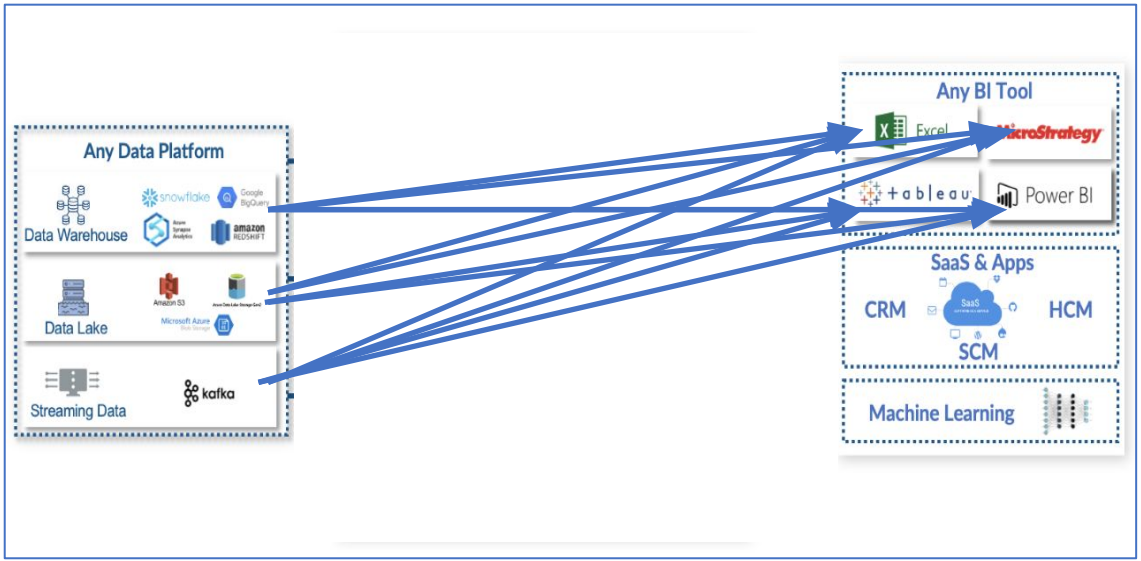


Data Storage Format Alternatives

- Cloud object storage is relatively inexpensive making data storage at high scale affordable
  - On-premises, specialized private cloud storage options such as Pure Flashblades tend to offer similar data type storage flexibility
- To take full advantage of the elasticity of the cloud without driving up costs, compute and storage need to be scaled separately
- To take full advantage of the many types of data available, such as Apache ORC, Apache Parquet, JSON, Apache Avro, etc., modern databases need to be able to analyze that data without moving or altering it
- A unified analytics warehouse that supports these various data formats means you have the benefit of querying them directly, without greatly expanding the hierarchical complex data types to a standard tabular data structure for analysis
- You should also be able to import data directly from these formats
- The ability to join data for analysis between the various internal and external data formats provides the highest level of analytic flexibility

Challenges of Today's Environment

- Complexity
- Context
- Cost

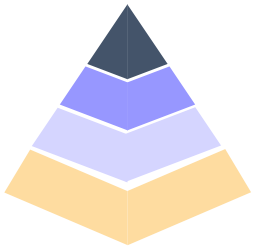


MCKNIGHT CONSULTING GROUP

Capabilities for Data Integration

About the Enterprise Contribution Ranking Report

- Focus is on value of product capabilities to the enterprise
- We believe AI including Generative AI is paramount to future-proofing technology and we take a keen focus on a solution's use of AI
- Assesses market leaders against critical capabilities of the market
- Focus is on generally available capabilities, but imminent realistic capabilities are included as well
- Expert-opinion ratings by our analyst-practitioners are used to rate each vendor against each capability, which nets out to an Advanced, Skilled, Partial or Beginner capabilities rating
- We furthermore plot each vendor on a quadrant based on axis of Project Scope Complexity and Project Technical Environment Complexity
- The most comprehensive and detailed ranking available
- Authored by hands-on Data Engineers
- The Enterprise Contribution Ranking Report is not sponsored; reprints are made possible



THE GORDIAN KNOT

The crux of efficient data utilization entails streamlined access, consolidation, governance, and adept management. Organizations need to continuously explore and comprehend new data sources, or fresh data within existing ones.

Enterprise Contribution Ranking Report: Data Integration

In alphabetical order, the 10 vendors chosen were:

1	Airbyte	6	Matillion
2	AWS: Glue	7	Microsoft: Azure Data Factory
3	Fivetran	8	Oracle
4	Google: Alooma	9	Qlik: Talend
5	Informatica	10	SAP

Comprehensive Native Connectivity and Multi-Latency Data Ingestion

- A strategy for easy interaction with various on-premises, multi-cloud, and SaaS data sources
- Utilizes specific data access APIs for richer and more adaptable services
- Key features:
  - Diverse data source onboarding: Handles change data capture, streaming data, batch processing, and real-time data ingestion
  - Automated data structure parsing: Adapts to changes in data formats without manual intervention
  - API-based connectivity: Enables microservices creation for richer and more adaptable services
  - Fast data source changes: Facilitates implementation of edge machine learning scoring models
  - Real-time and batch data import: Accommodates different data velocities
- Benefits:
  - Improved service richness, agility, and adaptability
  - More effective data ecosystem management
  - Enables machine learning and advanced analytics
  - Faster response to changing business requirements

API BASED CONNECTIVITY

Data Transformation

Data Integration Evolution	In the early stages of this market, we primarily relied on ETL (Extract, Transform, Load) for data transformation, which involves cleansing, enriching, and altering data during ingestion and delivery. However, in today's landscape, data integration needs to support a variety of integration patterns.
Versatility of data integration pipelines	Data integration pipelines should be versatile, accommodating both event-driven and bulk processing across batch, real-time, and streaming scenarios. In this setup, the entire data pipeline operates as a stream, capable of seamlessly incorporating data at rest.
Modern data integration tools	Modern data integration tools go further by offering templates, wizards, and even AI assistance to simplify the data integration development process, making it more accessible and productive

Data Security and Access Control

- Data masking
- Audit trails: Track access and modifications for compliance and troubleshooting
- Role-based access control (RBAC): Tailors access based on roles and data sets
- Encryption: Shields data in transit and at rest
- Authentication/authorization
- Data Loss Prevention (DLP): Proactively identifies and prevents unauthorized data exfiltration
- Implement consent management, anonymization, and data retention tools

Data Quality and Data Governance

This market acts as a single point of contact for the organization to source both internal and external syndicated data.

Applying data quality standards consistently throughout the entire organization is essential; labor-intensive manual techniques cannot be used in this capacity.

Data profiling is the first step in the process, giving information about the quality of the data before it is moved to a cloud data warehouse or data lake.

Quality control of data is critical no matter how or when it is transferred. An organization's data governance program provides the framework for upholding quality requirements.

Workflow Automation and Orchestration

- Workflow Automation:
  - Reduces manual intervention for repetitive tasks.
  - Ensures consistent and efficient execution.
- Orchestration:
  - Coordinates tasks and processes within a workflow.
  - Provides centralized control and visibility.
- Event-Driven Integration:
  - Triggers workflow adjustments based on data or system changes.
  - Ensures real-time responsiveness to new information.
- Scheduling:
  - Enables automated data processing at specific times or events.
  - Accommodates enterprise complexity with flexible options.

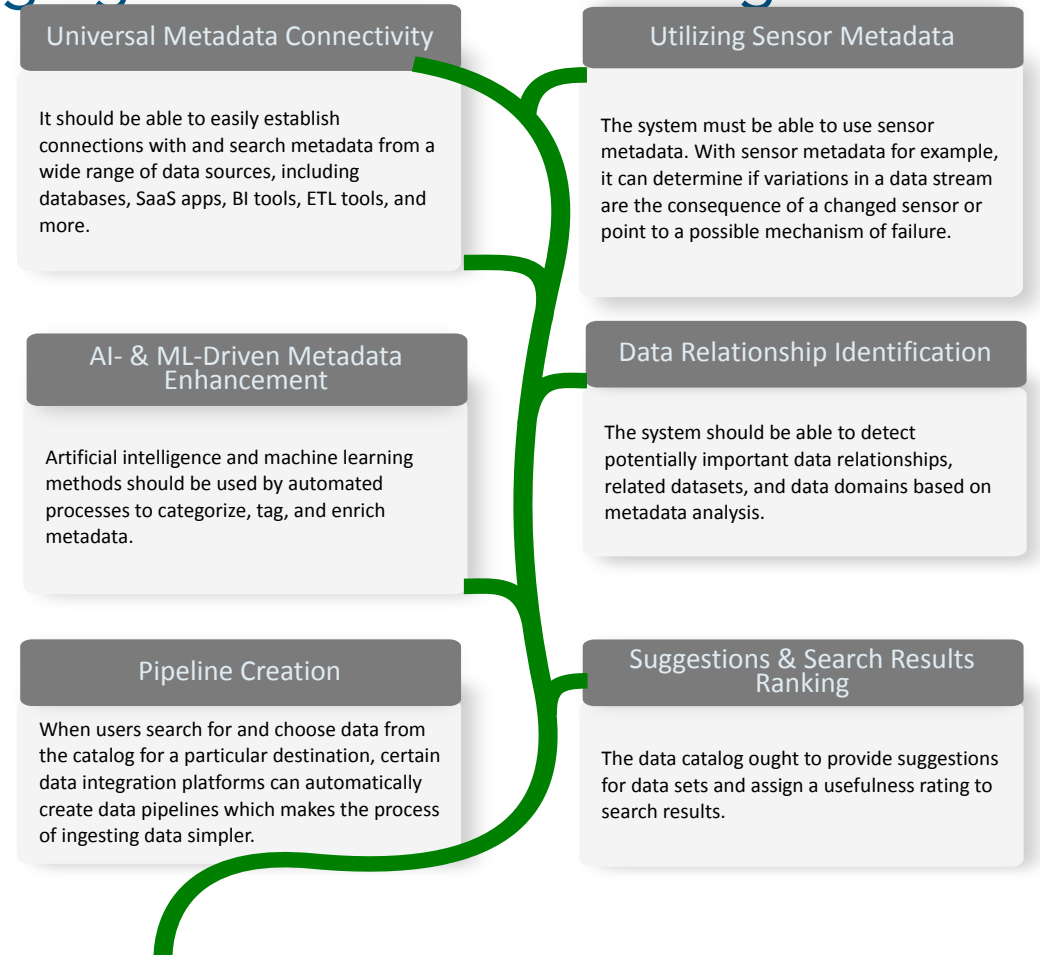
Analytics, Automation and AI

- Democratize Data-Driven Decisions: Facilitate user-friendly interfaces and functionalities for all levels of data exploration and analysis
- Future-Proof with AI: Leverage AI to automate data integration tasks and ensure your solution adapts to future advancements
- AI Advantage in Data Integration: Unlock the power of AI in data integration, cataloging, governance, and quality assurance processes
- Unlock Value Through Automation: Automate various tasks like data discovery, matching, categorization and lineage tracking with AI/ML
- Self-Healing Data Pipelines: Implement self-integration, self-healing, and self-tuning capabilities with AI-driven anomaly detection
- Prediction and Insights Made Easy: Integrate with machine learning frameworks for easier data-driven insights and prediction models
- Natural Language Processing: Leverage NLP to automate and enhance data integration processes by generating rules from text data

Seamless Integration with BI and Analytics Tools

- ✓ **Data Exploration and Reporting: Integration with BI tools empowers users to explore data, create insightful reports, and generate valuable visualizations**
- ✓ **Ad-Hoc Querying: Support for ad-hoc querying allows users to perform spontaneous and customized data searches, enabling quick responses to specific queries or analysis needs.**
- ✓ **Interactive Analysis: Interactive analysis capabilities enhance the user experience by providing dynamic, real-time data interactions and visualizations, fostering deeper insights into the data**

Data Cataloging and Metadata Management



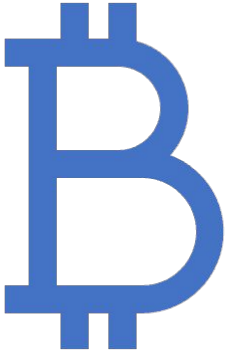
Enterprise Scaling with Performance

- The speed of syncing representative data sets between source platforms and target data platforms is essential selection criteria
- A fast-syncing process ensures that the target data platform stays up to date with the source platforms in real-time
- It allows for timely analysis and decision-making based on the most recent data available
- Without adequate speed, the entire pipeline will fail
- If the pipeline is successful, data volumes will grow
- Scalability is a critical aspect to consider
  - Your data integration platform must be able to scale horizontally or vertically to accommodate increasing data volumes and the evolving demands of systems and users
  - Performance optimization features such as query optimization and caching mechanisms are essential for delivering responsive and efficient services, particularly in professional services environments

Ecosystem Compatibility and Platform Versatility

- **Reliability**: The platform should be reliable and dependable, with minimal downtime or errors. This ensures consistent data availability and minimizes disruptions to business operations.
- **Ease of use**: The platform should be user-friendly and easy to manage, even for users with limited technical expertise. This includes intuitive interfaces, clear documentation, and readily available support.

Financial Operations, Compliance and Data Auditing



- Cost tracing and optimization functions are critical for effective cloud infrastructure cost management:
  - Track costs incurred and resources allocated to understand spending and optimize resource usage
  - Features like auto-scaling and resource allocation optimization help cut down on wasteful spending
- Robust monitoring and logging tools are essential for maintaining a well-running system:
  - Monitor system performance, track data flow, and identify issues proactively
  - Features like alerting mechanisms help administrators respond promptly to problems or deviations from expected norms
- Strong auditing capabilities and support for regulations like GDPR or HIPAA are vital for compliance and data integrity:
  - Monitor and record access to and modifications of data for accountability and transparency
  - Ensure data privacy and compliance with regulations through features like data encryption, access limits, and data anonymization

Evaluation of Enterprise Data Integration Capabilities

✓ Native Connectivity and Multi-Latency Data Ingestion 15%	✓ Analytics, Automation and AI 15%
✓ Data Transformation 5%	✓ Data Cataloging and Metadata Management 15%
✓ Data Security and Access Control 10%	✓ Enterprise Scaling with Performance 10%
✓ Data Quality and Data Governance 10%	✓ Ecosystem and Platform Versatility 5%
✓ Workflow Orchestration 5%	✓ Financial Operations, Compliance and Data Auditing 10%

Enterprise Contribution Ranking

Streaming Solutions

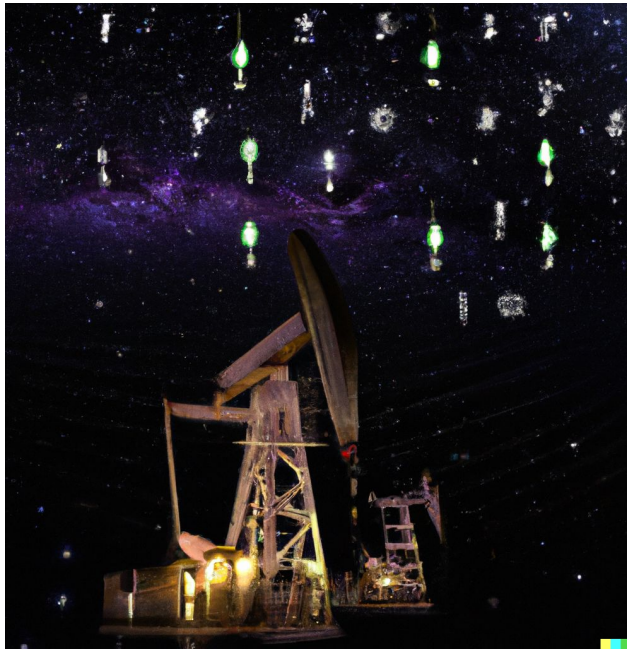
ETL is Insufficient for this combination

- Data platforms operating at an enterprise-wide scale
- A high variety of data sources
- Real-time/streaming data

- DI forces either real-time loading without being scalable or scalability with batch loading
  - Data, produced from numerous sources, is a torrent of flowing information, needing to be timestamped, dispatched, and even duplicated (to protect against data loss)
  - A postman is needed to distribute data from message senders to receivers at the right place at the right time.

Real-Time Data

- A.k.a. messaging, live feeds, real-time, event-driven
- Comes in continuously and often quickly
- Needs special attention and can be of immense value, but only if we are alerted in time
- Foundation for Artificial Intelligence
  - Stream data forms a core of data for artificial intelligence



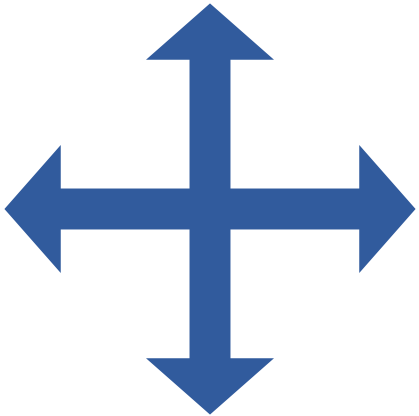
Enter Message-Oriented Middleware aka Streaming and message queuing technology

- Messages can be any kind of data wrapped in a neat package with a very simple header as a bow on top
- Messages are sent by "producers"—systems, sensors, or devices that generate the messages—toward a "broker."
- A broker does not process the messages, but instead routes them into queues according to the information enclosed in the message header or its own routing process
- Then "consumers" retrieve the messages from the queues to which they subscribe (although sometimes messages are pushed to consumers rather than pulled)
- The consumers open the messages and perform some kind of action on them

Performance and scalability in streaming

Throughput

High, sustainable rate of message processing



Storage

Ability to retain varying volumes of messages for varying lengths of time

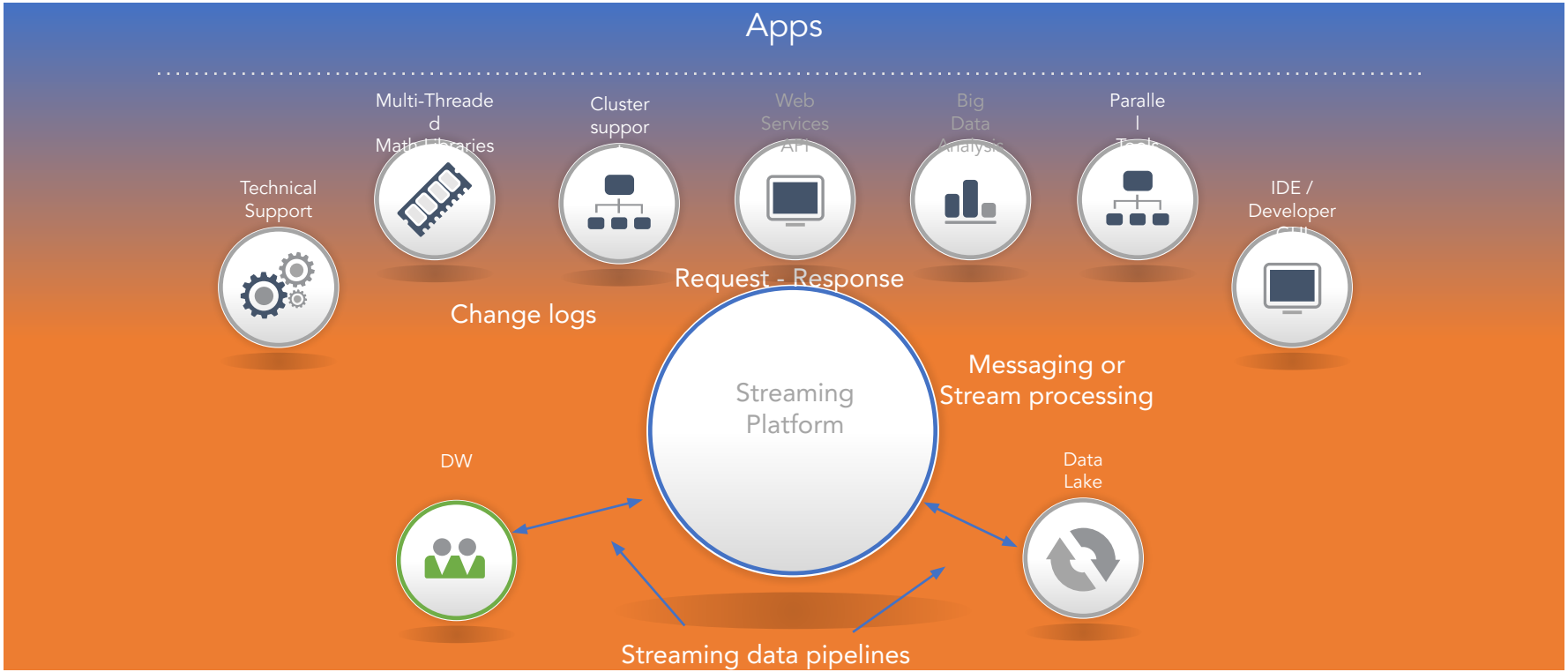
Latency

Fast, consistent responsiveness for publishing and consumption

Operations

Minimizing operational burden for scaling, tuning, and monitoring

Streaming Architecture



Apache Kafka

- Open source streaming platform developed at LinkedIn
- A distributed publish-subscribe messaging system that maintains feeds of messages called topics
  - Publishers write data to topics and subscribers read from topics
  - Kafka topics are partitioned and replicated across multiple nodes in your Hadoop cluster
- Enables “source to sink” data pipelines
- Kafka messages are simple, byte-long arrays that can store objects in virtually any format with a key attached to each message; often in JSON
- E&L in ETL through Kafka Connect API
- T in ETL through Kafka Streams API
- Fault-tolerant
- DIY

Apache Pulsar

- Originally developed at Yahoo
- Began its incubation at Apache in late 2016
- Has been in production as Yahoo for 3 years prior—utilized in popular services and applications like Yahoo! Mail, Finance, Sports, Flickr, Gemini Ads, and Sherpa
- Follows the publisher-subscriber model (pub-sub), and has the same producers, topics, and consumers as some of the aforementioned technologies
- Uses built-in multi-datacenter replication
- Architected for multi-tenancy and uses concepts of properties and namespaces
  - A property could represent all the messaging for a particular team, application, product vertical, etc.
  - Namespaces is the administrative unit where security, replication, configurations, and subscriptions are managed
  - At the topic level, messages are partitioned and managed by brokers using a user-defined routing policy—such as single, round robin, hash, or custom—thus granting further transparency and control in the process

Workloads are Distinguished by

- The number of topics
- The size of the messages being produced and consumed
- The number of subscriptions per topic
- The number of producers per topic
- The rate at which producers produce messages (per second)
- The size of the consumer's backlog (in gigabytes)
- The total duration of the test (in minutes)

Summary

- By necessity, we have numerous data stores for enterprise data
- Data integration is moving data; we want the tool to take care of the details
- Data Integration vendors show signs of growth that allowed them to drive customer-focused data strategies, make well-informed decisions, encourage innovation, and adjust to changing market conditions while gaining an advantage through optimized operations
- An enterprise's data stack would not be complete without data integration, and there are a wide variety of options available from vendors today to meet a wide range of requirements, skill levels, and use cases
- Message-Oriented Middleware aka Streaming and message queuing technology provide an Intelligent data platform for fast data
- An evaluator should consider whether they need a solution that is Full Spectrum, Solution Specific, Bespoke or Framework according to the intended application of data integration in terms of the complexity of project scope and how intricate the project's technical environment will be

MCKNIGHT
CONSULTING GROUP

Data Integration – Newsflash: We Still Just Move Data!

Presented by: William McKnight

"#1 Global Influencer in Big Data" Thinkers360

President, McKnight Consulting Group

3 X Inc 5000

/in/wmcknight

www.mcknightcg.com
(214) 514-1444