

Actian Data Platform Technical Overview and Architecture Guide

Technology and architecture
of the Actian Data Platform

Table of Contents

3	Executive Summary	
4	Overview	
7	System Architecture Overview	
9	Data Plane Architecture (Deep Dive)	
10	Maximizing Cloud Compute Resources	
11	Vectorized Execution and Exploiting Single Instruction, Multiple Data (SIMD)	
11	Utilizing CPU Cache as Execution Memory	
11	Storage	
11	Column-based Storage	
12	Automatic Storage Indexes	
12	Multi-tenant Hybrid Storage	
12	Real-time Update Capability (Our Secret Sauce)	
13	Data Compression	
13	Parallel Execution	
13	Data Integration	
14	High-Speed and Hybrid Data Ingestion	
14	Data Integration Design	
14	Data Connectivity	
14	Data Transformation	
15	Data Integration Management and Monitoring	
15	Data Quality	
15	Data Quality Design	
15	Data Profiling	
15	Data Remediation	
16	Data Quality Rules	
16	Data Quality Management and Monitoring	
16	Data Processing and Analytics Workload Support	
16	Data Lake Support	
17	SQL Support	
17	UDF support	
17	Security Framework	
17	Secure Storage and Encryption	
17	Network Security and Warehouse Isolation	
17	Security Maintenance and Compliance	
18	Handle Your Toughest Data Challenges with Confidence	
18	About Actian	

Executive Summary

Forward-looking business leaders must have trusted data-driven insights. Modern data platforms make this possible. They have the ability to integrate data while ensuring the quality needed for current and emerging use cases and to inform decision making at every level of the organization.

The Actian Data Platform offers a single, unified solution for real-time insights. The cloud-native data platform provides a comprehensive set of services to collect, manage, and analyze data. This includes data ingestion, storage, processing, analytics, and visualization capabilities. Built-in native integration makes it easy to load and transform data from external systems such as databases, streaming services, messaging platforms, and other sources.

From a business perspective, one primary reason to move to the cloud is to reduce cost, particularly with respect to shifting from a capital expense (CapEx) to an operating expense (OpEx). Yet a migration is about more than cutting costs; it's about achieving agility, speed, and scalability while minimizing risk. For example, the Actian Data Platform handles analytic workloads to meet the needs of a varied, demanding, and growing user base and complex datasets.

The Actian Data Platform, optimized for modern x86 CPUs, processes data faster than most other data warehouses and relational databases. In fact, a recent TPC-H benchmark test by McKnight Consulting Group found that the Actian Data Platform outperformed Google BigQuery and Snowflake with a performance of 11 times and three times faster than each vendor, respectively.

The Actian Data Platform is a fully managed cloud-native platform offering a comprehensive set of services to help organizations collect, manage, and analyze their data.



The benchmark also found that the Actian Data Platform achieved nearly eight times the performance of Databricks, over six times that of Snowflake, and an impressive 12 times the performance of BigQuery. Even with five concurrent users, Actian Data Platform maintained its performance advantage, outperforming Databricks by three times, Snowflake by over seven times, and BigQuery by 9.6 times. This faster performance enables data scientists and business analysts to iterate models, run real-time "what if" scenarios, and adapt quickly to changing business needs.

A fully managed platform built for modern organizations, the Actian Data Platform supports diverse analytics workloads. It offers a single platform to manage high-speed data ingestion, complex workloads, large data sets, and more users to deliver results fast and efficiently. This whitepaper details the technology architecture behind the Actian Data Platform.

The Evolution of Data Warehouses

Over the last five decades, data warehouses have evolved from static repositories to dynamic engines powering enterprise innovation. Today, they're more than a place to store data—they're pivotal for unlocking insights across vast, diverse, and ever-growing data sources. Data warehouses meet modern organizations' demands for real-time answers to complex queries, enabling both technical and business users to make critical decisions faster than ever.

Gone are the days of relying on IT for pre-packaged reports and dashboards. Today's users expect the freedom to explore, run ad-hoc queries, and discover insights on their own terms. This shift has pushed data warehouses from being rigid on-premises systems to becoming available in the cloud in hybrid environments for agility, scalability, and high performance to meet the demands of a rapidly changing data landscape.

Figure 1: Actian Data Platform Overview



Actian Data Platform – Overview

In today's data-driven landscape, organizations face the complex challenge of managing, integrating, and analyzing vast amounts of data from diverse sources. The Actian Data Platform is a fully-managed service that addresses this challenge by offering a unified solution that combines robust data integration capabilities with high-performance data warehousing, all within a single platform.

Actian has supported and provided operational services for mission critical database deployments for decades. Building on that experience, the Actian Data Platform combines the skills, scripts, tools, and best practices that our services and support staff, as well as our cloud operations staff, have built up over those decades. We've brought them together with our industry leading technologies to deliver a fully managed, cloud-native data platform that can collect, manage, and analyze data from edge to cloud.

Data Integration: Bridging Diverse Data Sources

The Actian Data Platform begins by seamlessly connecting to a wide variety of data sources, whether on-premises or in the cloud. It supports structured, semi-structured, and streaming data, ensuring that no matter the format or origin, data can be ingested and made ready for analysis. Key capabilities include:

- **Data Connectors:** Facilitates easy connection to an array of sources, from SaaS applications like ERP and CRM systems to IoT devices and social media platforms.
- **Data Ingestion (ETL/ELT):** Efficiently extracts, transforms, and loads data into the platform, ensuring that raw data is cleaned and transformed for optimal use.
- **Data Transformation:** Allows for the modification and structuring of data to meet specific analytical needs, providing the flexibility required for diverse use cases.
- **Data Quality:** Ensures the integrity and reliability of data, so that the insights generated are based on accurate and trustworthy information.

Data Warehousing: Powering Analytics and Insights

Once data is integrated, the Actian Data Platform’s data warehousing capabilities take over. This segment of the platform is designed to handle high-performance queries, store data elastically, and provide robust security and resilience. Key features include:

- **Query Engine:** Delivers fast, efficient data retrieval, enabling users to run complex queries and receive insights in real-time.
- **Elastic Cloud Storage:** Offers scalable storage solutions that adapt to organizations’ needs, ensuring they never run out of space while keeping costs manageable. The platform is cloud agnostic, allowing organizations to choose to store their data with the cloud service provider (CSP) of their choice.
- **Data Security:** Incorporates advanced security measures to protect sensitive data, aligning with best practices and compliance requirements.
- **Data Resilience:** Includes features like disaster recovery, ensuring that data is not only secure but also readily available in the face of unexpected disruptions.







Platform Services: Comprehensive Tools for Data Governance and Management

Cross-cutting these core capabilities is a layer of platform-wide Common Services that enhance the overall functionality and user experience of the Actian Data Platform. These services include:

- **API Services:** Enable seamless integration with other tools and platforms, enhancing the flexibility and extensibility of the platform.
- **Dashboards:** Provide visual insights and monitoring tools, making it easier for users to manage and analyze their warehouses’ and integrations’ usage and performance data.
- **Data Governance and Observability:** Ensure that data is stored responsibly, complies with regulations, and remains fully auditable, giving organizations confidence in their data management practices.
- **User and Policies Management:** Allows for the configuration of user roles and policies, ensuring that the right people have access to the right data.

Actian Data Platform is Built for Modern Data Needs

At its core, the Actian Data Platform is built on a set of key attributes that make it stand out.

Attribute	Description
 Distributed Architecture (MPP)	Leverages a massively parallel processing architecture for fast and efficient data processing.
 Elastic Compute (Kubernetes)	Uses container orchestration for scalable compute resources, ensuring that performance is optimized for any workload.
 Multi-Cloud Deployment	Provides the flexibility to deploy across multiple cloud environments, offering true hybrid cloud capabilities.
 Robust Disaster Recovery	Ensures that data is protected and can be recovered quickly in the event of a disaster.
 High Concurrency and Availability	Supports multiple users and applications simultaneously, with minimal downtime.
 Multi-Tenant Encryption	Protects data with advanced encryption techniques, even in multi-tenant environments.



Supporting Diverse Use Cases and Delivering Tangible Outcomes

The Actian Data Platform is designed to meet the varied needs of modern organizations, offering both high-level outcomes and specific business use cases.

High-Level Outcomes

- **Insights and Predictions:** Empower decision-makers with data-driven insights and predictive analytics.
- **Monetization:** Transform data into a revenue-generating asset by creating new data products, dashboards, models, APIs, and custom applications that can be offered to customers or internal stakeholders.
- **Data Products:** Develop and manage valuable data assets that can be utilized across various departments, enhancing operational efficiency and supporting strategic initiatives.

Business Use Cases

- **Operational Efficiency:** Organizations seeking to streamline their operations can leverage the platform to optimize workflows, reduce costs, and enhance overall productivity by centralizing and automating data processes.
- **Customer 360 Analytics/Improving Customer Experience:** The platform enables businesses to gain a holistic view of their customers, providing the insights needed to enhance customer experiences, personalize interactions, and build stronger customer relationships.

- **Regulatory Compliance:** For industries where compliance is critical, the platform offers robust governance and security features that help ensure adherence to regulatory requirements, reducing the risk of fines and reputational damage.
- **Risk Management:** By providing comprehensive data analysis and reporting tools, the platform aids in identifying, assessing, and mitigating risks, helping organizations to operate more securely and confidently.

The Actian Data Platform is more than just a data management solution; it's a comprehensive ecosystem that brings together data integration and data warehousing to empower organizations to unlock the full potential of their data. Whether the focus is on real-time analytics, data science, or operational systems, the Actian Data Platform provides the tools and capabilities needed to drive meaningful outcomes.

The Actian Data Platform Offers:

- One platform for all organization's data
- Real-time analytics
- Reporting and business intelligence
- Data monetization



Figure 2: Actian Data Platform – System Architecture

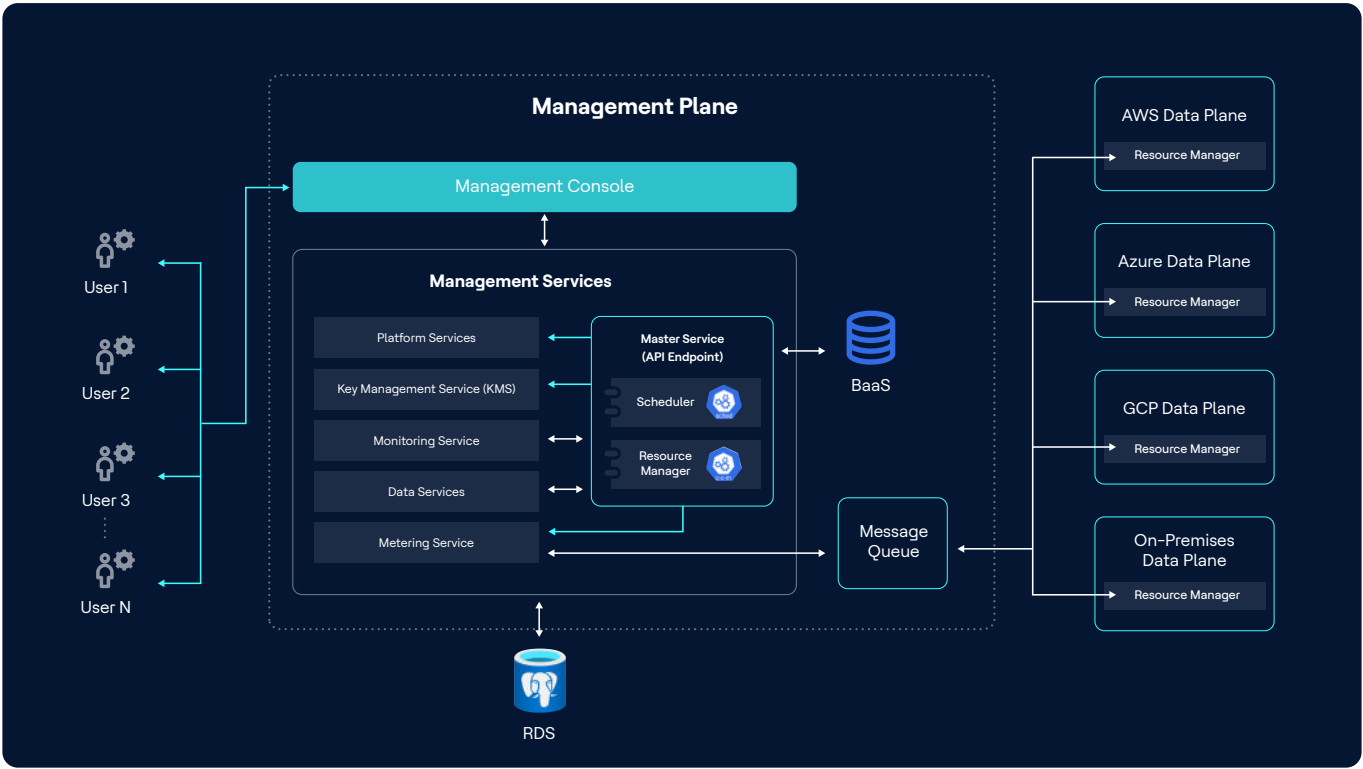


Figure 2 provides a technical breakdown of the different components and services within Actian Data Platform’s system architecture. Our architecture is designed for maximum compatibility with varying workloads and native integration with external data sources that ease keeping most data-related tasks under a single platform. Actian’s embrace of open standards prevents vendor lock-in, while maximizing compatibility with existing BI and analytics payloads.

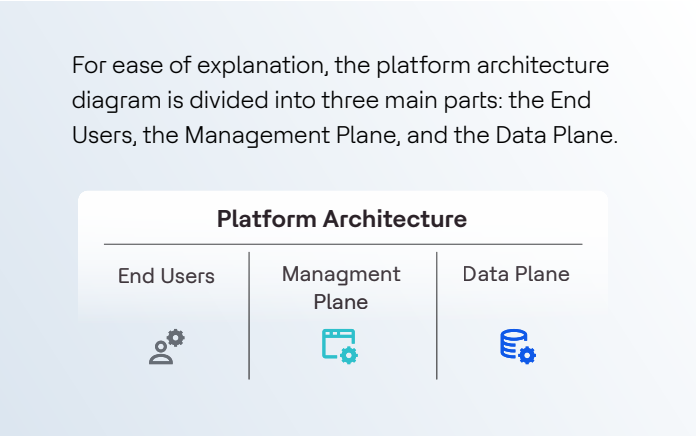
End Users

End users interact with the platform primarily through the Management Console within the Management Plane. This interaction is limited to high-level tasks such as starting, stopping, or creating new data warehouses within the platform. For operations that require direct access to the data within an existing data warehouse, users connect directly to the Data Plane, which handles all data-related activities.

Management Plane

The Management Plane is the control hub of the platform, responsible for overseeing the creation and management of data warehouses. It consists of three main components:

- 1. **Management Console:** This is the user interface layer that sits atop the Management Services. It provides users with an intuitive UI to perform management tasks. The Management Console communicates with the underlying services through a well-defined API, allowing users to issue commands and manage their data platform environment.

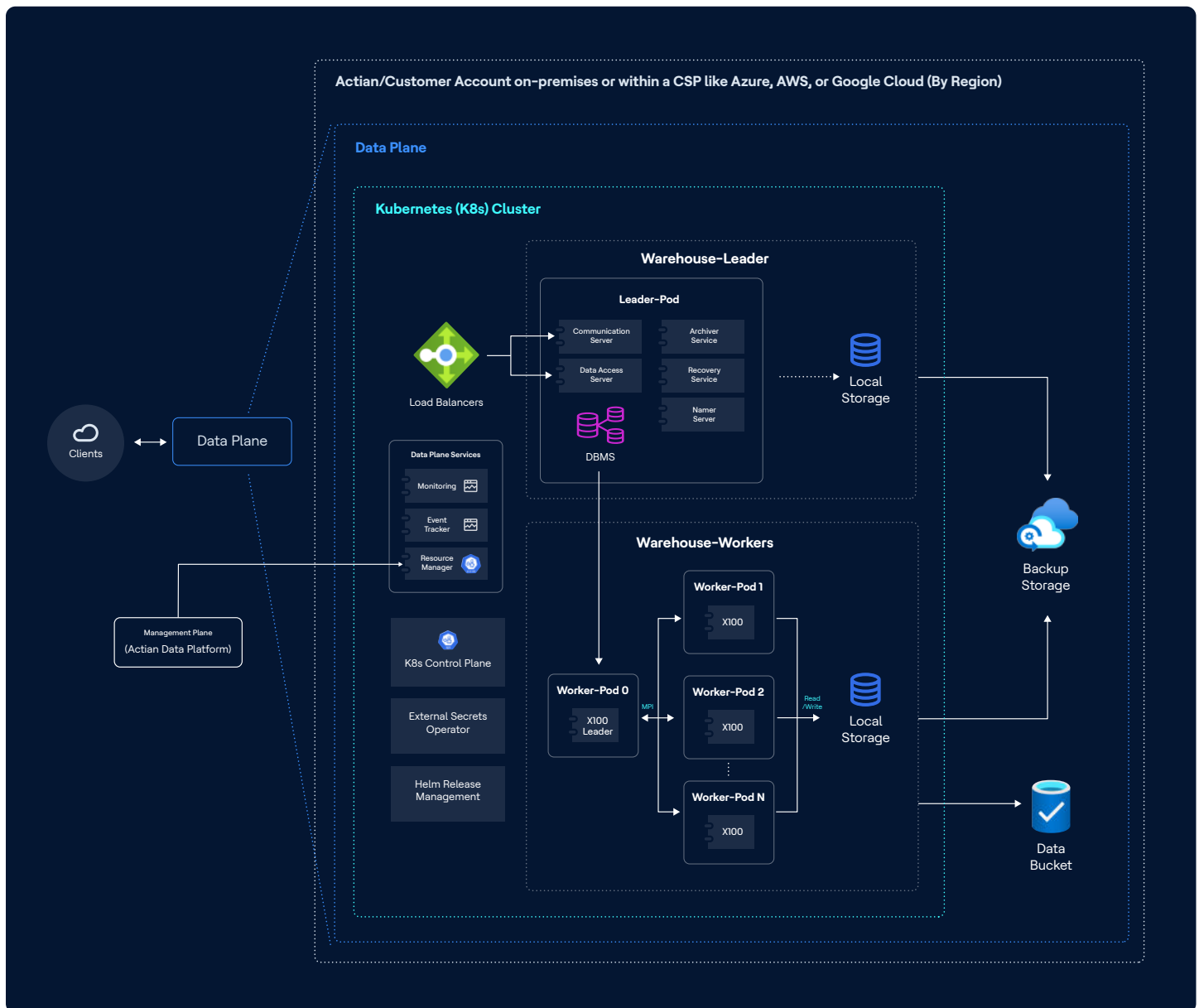


2. Management Services: This is a collection of micro services that coordinate various tasks within the platform. It is responsible for performing management and monitoring of compute / storage in the data plane. The key component here is the Master Service, an API endpoint that orchestrates the actions of other services within the Management Plane. Among these services, the "Data Services" group is particularly noteworthy, as it includes Integration Design, Data Profiler, and Integration Manager services—each playing a vital role in data integration and management tasks. The microservices utilize Amazon RDS and a homegrown Database as a Service (BaaS) for storing metadata, enabling easy scalability of databases as the demand and complexity of microservices increase. The key components include:

- **Master Service (API endpoint):** A critical service within the platform that handles logging, secure internet communication, and interactions with messaging and storage components. It plays a crucial role in managing warehouses within the data plane. It acts as the orchestrator for all API driven operations, and integrates with encryption and key management services. It is also responsible for user management and RBAC for warehouse operations.
 - **Scheduler:** Manages automation and scheduling operations within the platform. It ensures proper authorization for user-created tasks and is also responsible for running system operations such as backups and warehouse updates.
 - **Resource Manager:** The role of the Resource Manager within the Master Service involves updating both the Master Service (API endpoint) and the Resource Manager in the Data Plane using a Continuous Deployment (CD) tool, with responsibilities including syncing changes for the Resource Manager (Data Plane) across different cloud regions. This sync process involves following specific procedures to update cluster applications and ensuring the successful update of the Data Plane. These updates are part of the process where warehouse and database versions are controlled through the Master Service.
- **Platform Services:** A collection of services within our platform ecosystem that includes various endpoints and functionalities, such as authentication, API documentation, configuration, entitlement management, and upcoming metering features. It involves services like identity management, authentication federation, and managing essential instances such as the authentication and configuration services.
- **Action Key Management Service (AKMS):** A microservice that enables Customer-Managed Keys (CMK) functionality, acting as an intermediary for at-rest data encryption. It manages the interaction between the customer's Key Management Service (KMS) and the encrypted Action Data Platform's Warehouses' DataKey. The AKMS interacts with cloud providers' KMS as a backend service for AKMS, managing Customer's Master Keys for encryption and decryption processes.
- **Monitoring Service:** A collection of microservices responsible for managing and monitoring compute/storage in the Data Plane.
- **Data Services:**
 - **Integration Design Service:** Provides APIs for operations on integrations, connections, and services.
 - **Data Profiler Service:** Allows users to profile data, identify quality issues, and perform various data quality tasks. It involves creating data profiles, defining rules for data inspection, setting target connections for output, and reviewing results to enhance data quality. The service aims to understand user data quality processes, improve usability, and gather feedback for enhancements.
 - **Integration Manager Service:** Provides services for authentication, account and user management, job configuration, execution, and history tracking. Users can configure, schedule and monitor integration jobs, manage remote agents, upload files, and select runtime destinations.
- **Metering Service:** A microservice designed to collect, process, and report data on resource usage within the platform, enabling billing and analytics. It validates incoming data, processes it for billing purposes, and ensures data consistency and accuracy. The service handles multiple dimensions of metering, such as storage and compute, per region, and cloud provider, aiming to be scalable, adaptable, cloud-agnostic, secure, and highly available.

3. Message Queue: Serving as a messaging broker, the Message Queue facilitates communication between the Management Plane and the Data Planes. It interacts with the Resource Manager within the Data Plane, ensuring smooth transmission of commands and data.

Figure 3: Actian Data Platform – Data Plane Architecture



Data Plane Architecture (Deep Dive)

The Data Plane is the core environment where the data warehouses and corresponding storage reside. It hosts both the organization's data and the compute resources required to process that data. A Data Plane is a dedicated Kubernetes (K8s) cluster within a cloud account hosted by a CSP like Azure, Google Cloud, or AWS. This cloud account is where the main data buckets, along with any associated backups, are securely stored.

A Data Plane can be deployed to various locations, such as customer accounts in public clouds, enabling flexible deployment options. Security measures involve securing requests, communication, and access policies, with restricted access through IP allow lists and encryption. Additionally, the Data Plane's uptime and availability are critical, with services like the Master Service in Management Plane requiring specific availability links for testing.

- **Data Plane Services:** The data plane, which operates within Kubernetes, comprises a range of microservices. This is where the core resources of the Actian Data Platform, such as warehouses, databases, and integration runtime destinations are located. The dataplane services manage the lifecycle of Actian Data Platform resources and their corresponding data. Additionally, they use secure communication protocols when communicating with the management-plane services. These services are responsible for various tasks including resource scheduling for warehouses and databases, as well as overseeing real-time monitoring tools to track both the Actian Data Platform resources and their utilization effectively.
 - **Resource Manager:** A key component of the Data Plane is the Resource Manager, which manages the lifecycle of warehouse and database instances, facilitates message handling, and ensures efficient operation across the Data Plane.
 - **Monitoring:** The monitoring component of the Data Plane tracks warehouse and database usage and activities, facilitating operations like auto-starting and automatic sleep or stopping when idle.
 - **Event Tracker:** The event tracker serves as a valuable tool for efficiently managing and monitoring Actian Data Platform resources by capturing Kubernetes events. These events are then relayed to the management plane to facilitate effective resource status management. Additionally, the event tracker plays a crucial role in identifying possible resource issues by analyzing recurring event patterns.
- **External Secrets Operator:** To share sensitive data between the Management Plane and Data Plane services, the Kubernetes External Secret Operator is utilized. This operator helps in securely managing and distributing secrets within the Kubernetes environment.
- **Helm Charts Release Manager:** It facilitates the deployment of Actian Data Platform resources within Kubernetes clusters, ensuring enhanced reliability and consistency in resource deployments.
- **Warehouse-Leader:** The warehouse leader node ensures fault tolerance by monitoring the health of worker nodes and resource utilization within the warehouse cluster. It manages client connections and processes incoming queries. Query processing involves parsing, optimizing, and scheduling execution within the warehouse.
- **Warehouse-Workers:**
 - **X100 Leader:** The leader component in Warehouse workers is responsible for coordinating and managing query execution across the distributed system. The X100 leader component plays a crucial role in optimizing query performance and redistributing tasks as needed.
 - **X100 Worker:** Warehouse X100 worker nodes process queries and perform operations on stored data, including retrieval, aggregation, filtering, and sorting. These nodes are vital for parallelizing workloads to enhance database system performance and scalability by collaborating with other warehouse nodes.
- **Cloud Storage:**
 - **Warehouse Data:** An organization's data is securely stored in dedicated cloud storage (Shown as "Data Bucket" in the diagram), offering numerous benefits such as scalability, cost-effectiveness, data retention, high durability to prevent data loss, and compliance with industry regulations.
 - **Backup:** The warehouse backups are stored in separate cloud storage during the weekly maintenance window. This approach offers benefits such as warehouse restore, cloning, and disaster recovery, enhancing data protection. Storing data in cloud storage buckets ensures consistent and reliable backup processes for the user's data management needs.

Actian Data Platform – Maximizing Cloud Compute Resources

The Actian Data Platform was developed from the ground up to take advantage of performance features in commodity CPUs, resulting in dramatically higher data processing rates compared to other analytic solutions. It's based on tried and true architectural design and operational experience cultivated from decades of database technology development, deployment, and management.

The Actian Data Platform is able to accelerate compute resources by taking advantage of powerful CPU features that most other data warehousing solutions don't, accelerating on-premises and cloud environments alike.



Examples include so-called SIMD instructions, larger chip caches, super-scalar functions, out-of-order execution, and hardware-accelerated string-based operations. In fact, most of today's analytics software that was originally written between the 1970s and mid-90s has become so complex that, to take advantage of these performance features, a complete rewrite of the technology would be required. More recent Open Source Big Data technologies, while written for the latest hardware, are often sub-optimized and unproven at scale.

Vectorized Execution and Exploiting Single Instruction, Multiple Data (SIMD)

SIMD enables a single operation to be applied to a set of data at once. The Actian Data Platform takes advantage of SIMD instructions as well as Advances Vector Instructions (AVX) by processing vectors of data through the respective instruction sets where possible. Because typical data analysis queries process large volumes of data, using SIMD and AVX may result in the average computation against a single data value taking less than a single CPU cycle.

At the CPU level, traditional databases process data one tuple at a time, spending most CPU time on overhead to manage tuples and not on the actual processing.

In contrast, the Actian Data Platform processes vectors of hundreds or thousands of elements at once, which effectively eliminates these overheads. In addition it provides opportunities to apply SIMD instructions as each element in a vector requires the same operation being executed. As a result, CPU resources are used to maximum effect to perform the actual work.

Utilizing CPU Cache as Execution Memory

Most improvements to server memory (RAM) over the last few years have resulted in much larger memory pools but not necessarily faster memory access. As a result, relative to the ever-increasing clock speed of the CPU, memory access has become slower over time. In addition, with more CPU cores requiring access to the shared memory pool, contention can be a bottleneck to data processing performance.

To achieve maximum data processing performance, the Actian Data Platform avoids using shared RAM as execution memory. Instead, the platform uses the private CPU core and CPU caches as execution memory, delivering significantly greater data processing throughput. Vectorized execution ensures that data being processed fits the CPU caches and as many instructions as possible are executed against the data before leaving the cache again.

Action Data Platform – Storage

The Actian Data Platform uses cloud storage—and all the resiliency benefits it brings—to keep data safe. Data volumes are encrypted for security. The data itself is partitioned horizontally to optimize performance. The platform separates storage and compute resources, enabling compute resources to be shut off when not in use. This means organizations only pay for data storage when the data is not in use, saving costly compute costs for when data is being actively processed or analyzed.

Column-based Storage

When relational database software was first written, it implemented so-called row-based storage: all data values for a row were stored together in a data block. Data was always retrieved row by row, even if a query only accessed a subset of the columns in a row. This row-based storage model works well for On-Line Transaction Processing (OLTP) systems in which the data stored is highly normalized, tables are relatively narrow, queries often retrieve very few rows, and many small transactions need to be processed.

Why row-based storage is subpar for On-Line Analytical Processing (OLAP) systems:

- Tables are often (partially) denormalized, resulting in many more columns per table. Typically, not all columns are accessed by queries.
- Most queries access data from many rows, but result sets are typically small.
- Row-based storage cannot easily compress data at rest as table rows typically hold heterogeneous data compared to more homogeneous data in a single column.

As a result of these differences, a row-based storage model typically generates a lot of unnecessary I/O for a data warehouse workload. A column-based storage model, in which data is stored together in data blocks on a column-by-column basis, is generally accepted as a superior storage model for data analysis queries.

The advantages are:

- Only fetch requested columns, and the related data, from disk.
- Achieve high compression rates at low compression/decompression cost optimizing I/O and bufferpool space as the platform buffers compressed blocks. See section on data compression.
- Denormalized data, where columns hold repeated values, particularly benefit from column wise compression, effectively hiding the redundancy introduced by denormalization.
- No pollution of bufferpools with data not required for processing in case of repeated queries on the same data.

In OLAP setups data is added through a controlled rather than an ad-hoc process, and often large data sets are added at once or through an ongoing (controlled) stream of data. In many cases data is only "added" and not "modified." This helps with the Achilles' heel of columnar storage. Where an OLTP insert/update typically requires a single disk block update, in columnar storage a block per column needs to be updated. Doing that at the end of a table is a manageable overhead, updating data in the middle of the table requires special care. See section on real time updates.

Automatic Storage Indexes

The Actian Data Platform automatically maintains a storage index per column, storing minimum, maximum, and a sample value for each data block. The storage index is very efficient in determining whether a database block is a candidate block for a particular query either because of explicit filter criteria or implicitly as a result of processing table joins. Based on these indexes a typical OLAP query will only access a fraction of the full data set.

Consider the example of a frequently updated table with new data rows. If new data holds a date/time information of creation, effectively having a growing domain for the entries, then automatic storage indexes will very efficiently support the pruning of blocks on request like "Generate a list of items sold last week."

Multi-tenant Hybrid Storage

Hybrid storage environments fulfill a crucial need that previous generations of cloud data warehouses often couldn't: the ability to store sensitive data on-premises while providing the same managed service experience, technology compatibility, and application support as their cloud counterparts. The multi-tenant storage capability is particularly important for industries like financial services, healthcare, and pharmaceuticals, where regulatory compliance demands stringent data management and analytics for sensitive information.

These enterprises want to leverage the same technologies for their analytics needs to write applications that seamlessly join data that resides on-premises and in the cloud. The Actian Data Platform is a component in a broader cloud strategy, supporting integration with hundreds of data sources including Oracle and SAP as well as popular SaaS solutions like Salesforce, NetSuite, Workday, and ServiceNow. Data from these services can be imported at a user-owned and user-defined cadence and seamlessly blended to provide a 360-degree view.

Real-time Update Capability (Our Secret Sauce)

A big challenge with most column-based databases is incremental small inserts, updates, or deletes (as opposed to large bulk data load operations at the end of a table). The Actian Data Platform meets this challenge with high-performance in-memory Positional Delta Trees (PDTs). The Actian platform uses PDTs to store small incremental changes, as well as updates and deletes.

A PDT is an in-memory structure that stores the position and the change (delta) at that position. Queries efficiently merge the changes in PDTs with data stored on disk live at scan time. Because of the in-memory nature of PDTs, small DML statements can be processed very efficiently. A background process writes the in-memory changes to disk once a memory threshold is exceeded.

The Actian Data Platform implements an ACID2-compliant transactional database with multi-version read consistency. It ensures visibility of all previously committed transactions, including incremental and bulk data loads. For recoverability, changes are persistently written to a transaction log before commit completion. PDTs enhance efficiency in conflict detection and accelerate point inserts/updates compared to standard delta update structures like heaps.

Data Compression

The Actian Data Platform compresses data on a column-by-column, page-by-page basis using any of the following algorithms or a combination of them:

- **Run-Length Encoding (RLE):** A data value is stored as well as the number of subsequent values that are the same. This compression algorithm is very efficient on ordered data with relatively few unique values.
- **Patched Frame Of Reference (PFOR):** A base value is determined per data block, and other values in the same block are encoded by storing the difference with the stored value using as few bits as possible. This is beneficial because the range of the actual data is typically much smaller than the range of a used datatype. What makes PFOR special compared to similar solutions found in other products is the treatment of outliers. For example, if 99% of values are in the range 0–255, and 1% of the values is very large (e.g., around a million), then with PFOR the majority of the data will be stored using only one byte, while other solutions would use 2.5 bytes.
- **Delta encoding on top of PFOR:** To reduce the values of the integers with PFOR, it is sometimes more efficient to store the delta from the previous value and then apply PFOR compression. This can be very efficient on ordered data.
- **Dictionary encoding:** This method stores pointers to a dictionary of unique values. This algorithm is very efficient for a limited number of very frequently occurring values. Similar to PFOR, the used encoding also allows for outliers to be handled separately.

- **LZ4:** This method detects and encodes common fragments of different string values. It is particularly efficient for medium and long strings where dictionary encoding fails to deliver good results.

The algorithms used by the Actian platform to compress data were selected for their speed of decompression over a maximum compression ratio. The compression ratio organizations can achieve with the Actian Data Platform is highly data dependent: 4–6x compression ratios are common for real-world data but both lower and higher compression ratios have been observed. This balances CPU usage and bandwidth usage to ensure maximum processing throughput.

Parallel Execution

The Actian Data Platform implements a flexible adaptive parallel execution algorithm. The platform executes statements in parallel using any number of CPU cores and will intelligently balance concurrency and query parallelism.



Actian Data Platform – Data Integration

The Actian Data Platform delivers flexibility that empowers reliable transactions and trusted, real-time analytics, making it easier to get from data source to decision with confidence. The platform is designed to make data easy by offering services that address the challenges of data integration and data quality.

Data integration, offered as a service on the Actian Data Platform, goes beyond traditional integration tools. Modern capabilities enable organizations to access and ingest data. Access and ingest data from any system, application, or resource, both on-premises and in the cloud. The platform enables organizations to:

- Reuse integrations, allowing users to develop integrations once and deploy anywhere
- Integrate without IT resources by using codeless, low-code, and pro-code integration options
- Run integrations natively on the Actian Data Platform in the cloud or in a hybrid environment
- Manage and monitor all Integration and data quality jobs in one environment whether running in the cloud or in hybrid environments

High-Speed and Hybrid Data Ingestion

Action Data Platform customers use a wide variety of methods for loading the data warehouse. Some use traditional ETL and ELT tools; some use the native bulk-loader or the desktop file-loader to import local data files; others prefer to use Spark for high-speed ingestion of large volumes of data.

Organizations can perform traditional ETL operations including data transformation, data quality, and data cleansing in parallel, eliminating traditional performance bottlenecks in the data acquisition processes. Moreover, organizations can use Action DataFlow to perform real-time analysis of this data-in-motion, looking for predetermined patterns and/or outliers to business models, and take prescribed action when these patterns are observed.

Data Integration Design

Action Integration includes two design options. The first is a native designer on the Action Data Platform targeting data analysts. This designer provides the ability to edit schemas and transform data in an easy-to-use tooling. The second option is an advanced designer (DataConnect) for data engineers to design complex schemas, transformations, data quality rules, and data pipelines which can be deployed on-premises, in the cloud, or in hybrid environments.

Action Integration offers comprehensive data pipeline capabilities encompassing orchestration, scheduling, and management of data pipelines, ensuring users access accurate and comprehensive data promptly, whether in batch, real-time event, or embedded modes. All integrations can be run on the Action Data Platform with full management and monitoring capabilities.

Data Connectivity

Action Integration offers an extensive library of over 200 pre-configured data connectors, facilitating seamless access to a range of cloud and on-premises systems, legacy platforms, third-party application databases, and leading BI and analytics tools. Furthermore, the Action Data Platform empowers users to swiftly develop custom connectors tailored to cloud-based data sources with REST and SOAP APIs.

When leveraging the native design capabilities of the Action Data Platform, there is a limited set of pre-built connectors that can be used as sources for integrations. There is also the ability to easily and quickly create source connectors to any SOAP or REST based APIs in the cloud. This makes the number of source connectivity almost limitless. On the target side of the connectivity story, there is only one: the Action Data Warehouse. Note: this is planned to change in the future.

When leveraging the DataConnect Design Studio, there are more than 200 connectors that can be used as sources or targets within an integration job. These sources and targets can be on-premises or in the cloud and are designed to easily transform bulk quantities of data in real-time. When leveraging DataConnect on the Action Data Platform, agents are used to connect to any on-premises sources or targets. Agents on the platform are managing and monitoring instances of DataConnect on-premises which pass data to/from the platform.

Data Transformation

The Action Data Platform provides robust transformation capabilities which can be run in any environment: cloud, on-premises or hybrid.

Data analysts can transform data with an easy-to-use UI to map data to warehouses on the Action Data Platform. Transformation is done as part of an integration from any cloud source to the Action Data Warehouse. Transformations from flat, relational data will be with a single source and a single target. Transformations for JSON based source data (hierarchical) will have a single source and multiple targets. The transformation is built through the expression builder within the Create Integration wizard. The expression builder allows users to drag and drop expressions. These expressions will click together to form a larger expression. These can be tested within the expression builder to ensure correctness.

Data engineers can perform advanced transformations using the DataConnect Design Studio which can be deployed on-premises, in the cloud or in hybrid environments. DataConnect transformations can read data from multiple sources and write to multiple targets in any structured or semi-structured format including hierarchical data of any complexity such as JSON, XML or data complying with industry standards. Fully customized, advanced transformation rules are built leveraging the large library of in-built functions.

Data Integration Management and Monitoring

Action Data Platform includes dashboard functionality to manage and monitor all integrations. Features include:

- The platform provides a unified interface for managing all integration processes, making it easier to monitor, configure, and manage different data pipelines from a single point of control.
- Integrations can be automated and scheduled to run at specific times. This ensures that data is always up to date without the need for manual intervention.
- Actian provides real-time monitoring of data integration processes. Our dashboard shows all integration and data quality jobs running in the cloud or hybrid environments.
- Detailed logs and audit trails are maintained for all integration jobs. These logs provide insights into the execution of integration jobs, including information on errors, processing times, and data transformations.

Action Data Platform – Data Quality

To fully harness the value of data, it is crucial to ensure that the data is reliable. This involves validating the quality of attributes like completeness, accuracy, and timeliness. Organizations must gain an in-depth understanding of the data locations and data types across various systems to achieve this. Once this is achieved, the focus should shift towards integrating data from diverse systems and ensuring its accessibility for business intelligence, analytics, and application purposes.

To begin integrating data, it is essential to engage in strategic planning and identify necessary resources and technologies for data quality. Data should be continuously monitored as it moves between departments and applications. Providing users with high-quality data improves business outcomes and promotes a data-driven culture within the organization.

Data quality is the process of understanding, validating, and cleansing the data before moving to a data warehouse or application. This ensures that the data is in the required format and meets predetermined quality standards. Ultimately, prioritizing data quality ensures businesses have reliable data to make well-informed decisions and accelerate business growth.

The Actian Data Platform automates data verification and

remediation, helping organizations establish data quality standards. Data that doesn't adhere to quality rules can be separated from other data, avoiding data bottlenecks. This data can then be cleansed manually or through automated workflows, before it is integrated into the data warehouse or utilized by other applications. By implementing these measures, the Actian Data Platform ensures that data integrity is maintained, and any potential issues are addressed, before they impact the business.

Data Quality Design

Just like Actian Integration, Actian Data Quality includes two design options. The first is a native designer on the Actian Data Platform targeting data analysts. This designer provides the ability to understand data by creating Data Profile Rules in an easy-to-use tooling. These rules allow organizations to split their valid data from their invalid data. The second option is an advanced designer (DataConnect) for data engineers to design data profile rules and data remediation rules alongside the integration functionality. DataConnect jobs can be run in the cloud or in hybrid environments.

Data Profiling

Data profiling plays a pivotal role in this process, as it provides an understanding of each data set. This includes analyzing individual attributes, understanding relationships among data, and determining patterns and anomalies. By conducting data profiling, businesses can mitigate risks associated with inaccurate or inconsistent data. Successful implementation of data quality can greatly enhance the efficiency of pipeline creation and overall integration processes. It allows companies to identify and address any issues with the data, ensuring its reliability and usefulness for decision-making purposes. With proper data profiling, businesses can have confidence in the integrity of their data, leading to more accurate analysis and improved business outcomes.

Data Remediation

Data remediation refers to the process of identifying and correcting errors, inconsistencies, and inaccuracies in source data. This can include tasks such as removing duplicate records, standardizing format and data types, and filling in missing values. Understanding data abnormalities with Data Profiling is important, but being able to correct/remediate the incorrect data can be a core functionality for some organizations.

Data Quality Rules

There are two types of Data Quality rules that can be used within the Actian Data Platform today: profile rules and remediation rules. Data Profiles can be created natively on the Actian Data Platform or within the DataConnect designer. Data remediation rules can only be created within the DataConnect designer, but designing natively on the platform will be coming soon.

Profiling rules can help identify problems in source data. These rules are of the following types:

- Some rules can be used to generate aggregate statistics, which help identify inaccuracies by examining aggregated values over large datasets.
- Other rules are test rules that generate pass and fail statistics and also route pass and fail records to two different targets respectively.

As mentioned above, Remediation rules help to identify and correct errors in data. During Profile execution, the Remediation rules are processed first followed by the Profiling rules. The processing order ensures that the changes specified in the Remediation rules are applied to data before the Profiling rules are evaluated. Rules can be configured in any order. While organizations can rearrange the Remediation rules by moving them up or down, the Profiling rules cannot be rearranged. The Remediation rules are executed in the order organizations create them or rearrange them on the Rules tab and the results may vary as per their order, however, the Remove Duplicates and Remove Duplicates Fuzzy Matching rules will always be processed last.

Data Quality Management and Monitoring

Monitoring data integration and quality is essential to ensure accurate and reliable insights. Fluctuations in data can lead to a decrease in the overall quality of the information being gathered and integrated. Organizations can identify quality issues or data inconsistencies by maintaining ongoing oversight and taking corrective actions promptly. This proactive approach helps to prevent any negative impacts on decision-making processes that rely heavily on data analysis.

Actian Data Platform includes dashboard functionality to manage and monitor all data quality jobs. Features:

- The platform provides a unified interface for managing all data quality processes, making it easier to monitor, configure, and manage different data integration and data quality pipelines from a single point of control.
- The dashboard provides a historical view of their profile

executions with charts showing how the quality of the dataset has been performing over time. Under run history, there are options to drill down into column level and rule level which enables users to see a particular column's data quality over time.

- Data quality jobs can be automated and scheduled to run at specific times. This ensures that data is always up to date without the need for manual intervention.
- Actian provides real-time monitoring of data quality processes. This dashboard shows all integration and data quality jobs run in the cloud or hybrid environments.
- Detailed logs and audit trails are maintained for all data quality jobs. These logs provide insights into the execution of the jobs, including information on errors and processing times.

Actian Data Platform – Data Processing and Analytics Workload Support

The Actian Data Platform enhances analytic data processing through an ANSI-compliant relational database, delivering high performance previously achievable only with expensive on-premises data warehouses or complex, tuned systems. It serves organizations requiring a relational database with ANSI SQL support and industry-standard JDBC/ODBC interfaces, offering superior speed and cost-effectiveness compared to in-memory databases without memory limitations.

The platform's architecture addresses needs across various roles, from data architects to business analysts. It supports multiple programming languages and tools, including SQL, Python, Jupyter, R-Lib, TensorFlow, and BI tools like Microsoft PowerBI and Tableau.

Data Lake Support

The platform utilizes external tables to access data lake content, supporting data types convertible to internally supported formats. It processes Parquet, ORC, JSON, and CSV files on-the-fly. Open table formats, such as Iceberg, are accessible via external tables, enabling cross-system interoperability. The platform reads files from major cloud storage providers (S3, GCS, Azure). Data lake preprocessing, executed via SparkSQL and Scala programs, allows integration of any data lake item matching an internally supported data type. While the Actian Data Platform currently does not offer native data lake capabilities, it provides robust import functionality for various data types from data lake sources, ensuring efficient data integration and analysis within the platform ecosystem.

SQL Support

The Actian Data Platform complies with SQL 2016 standards, incorporating analytics capabilities like CUBE, ROLLUP, GROUPING SETS, and analytic windowing functions. This standards-based approach mitigates vendor lock-in, facilitating direct migration of existing workloads with performance improvements.

UDF Support

User-Defined Functions (UDFs) are supported in Python, JavaScript, SQL, Scala, and TensorFlow. The platform allows scalar UDFs, system extensions with custom functions, and DBA-implemented UDFs available to all users. Python UDFs enable ML scoring, ScalaUDFs leverage the Spark framework, and TensorFlow UDFs provide efficient model scoring.

Actian Data Platform – Security Framework

Public cloud security perception has evolved from initial skepticism to recognition of robust security measures. Cloud service providers now effectively combat threats through continuous diagnostics, monitoring, and mediation.

The Actian Data Platform's security posture resembles a multi-layered "onion" approach, with each cloud service incorporating multiple security layers, each monitored and alerted. Our service monitors our console webpages for vulnerabilities, with scans running each weekend and generating email reports. This continuous monitoring ensures prompt detection and addressing of potential security issues.

Building upon major cloud providers' infrastructure security, the Actian Data Platform Cloud Security Framework includes:

- Single and multi-tenant data management features
- White-listing and non-public facing IP/ports
- User-based, group-based and role-based access control
- Data encryption, including column-level data at rest encryption on top of the encrypted file-system
- Data masking
- Dynamic and ongoing security auditing and security alarms

Secure Storage and Encryption

The Actian Data Platform leverages underlying cloud provider security measures. For AWS, this includes AES 256-bit encryption for data at rest and FIPS 140-2 for key encryption during data transfer to/from S3, ADLS, Google Cloud Storage, and between Actian Data Platform instances or external customer repositories.

Internal storage is secured through file system encryption with block-level corruption detection and replacement. The platform offers industrial-grade encryption for data-at-rest, in-transit, and optional block-level encryption. Each warehouse is isolated using industry best practices, with no inter-warehouse communication.

Data loading occurs over HTTPS/TLS, with private channel transfer available in specific configurations (currently AWS-only). Google Cloud Storage employs server-side AES-256 encryption by default. Customer-supplied keys can manage block-level encryption within warehouses.

Network Security and Warehouse Isolation

The architecture comprises a management plane and multiple data planes, isolating management functionality from warehouse data ([learn more](#)). Data planes operate in dedicated VPCs with security group isolation. Istio service mesh provides an additional security layer, while Calico enforces strict ingress and egress rules.

Warehouses are fully isolated. Inbound warehouse traffic is routed through a public load balancer and restricted by IP allow lists.

Security Maintenance and Compliance

Actian Data Platform is hosted in ISO 27001 certified data centers with high-level physical security, including biometric access and 24/7 surveillance. Users can specify geographical regions for warehouse creation.

The platform provides database-level log events within warehouses and security event logs at the management plane level. Regular patching occurs within defined maintenance windows. The platform undergoes PCI-approved vendor scans and web application vulnerability assessments. A global operations team provides 24/7 monitoring, with performance and uptime metrics available via the Customer Portal and RSS feed. Current performance and uptime metrics are available at the [Actian Data Platform Customer Portal](#) and through RSS feed.

The bigger the data volumes, the more users on the system, the more complex the queries, the better the platform performs.



Handle The Toughest Data Challenges with Confidence

The Actian Data Platform is a fully managed cloud data warehouse environment, designed for hybrid deployment, that provides pre-integrated connectors to hundreds of popular data sources such as Salesforce, NetSuite, WorkDay, and ServiceNow.

The Actian Data Platform is engineered to handle the toughest data, user and query volumes for massive scalability. The bigger the data volumes, the more users on the system, the more complex the queries, the better the platform performs. Getting started with the Actian Data Platform is easy and requires just three steps:

1. Create the Actian Data Platform cluster
2. Load the data
3. Start querying

The production cluster could be up and serving the business needs in as little as 20 minutes.

The Actian Data Platform empowers enterprises to make the right decisions by:

- Accelerating business intelligence for the entire organization
- Effortlessly supporting hundreds of concurrent users
- Eliminating stale data through continuous real-time updates

The Actian Data Platform provides a single platform for business analysts, developers, data scientists, and data engineers. Actian enables business to:

- Run advanced data analytics at scale with sub-second performance
- Analyze data from terabytes to petabytes
- Breeze through both mixed and complex analytic query workloads

With Actian, your organizations will foster faster innovation. Now, organizations can shorten AI and machine learning life cycles with parallelized data loads. Because of the performance it delivers, no sampling is required. And Actian Data Platform's open architecture integrates with R, Python, Spark, and dozens of other tools.

There is nothing as compelling as seeing an organization's own queries running against their own data. Prepare to be amazed. Get started with the Actian Data Platform today, for free. Learn more at www.actian.com/demo-request.

About Actian

Actian makes data easy. We deliver cloud, hybrid, and on-premises data solutions that simplify how people connect, manage, and analyze data. We transform business by enabling customers to make confident, data-driven decisions that accelerate their organization's growth. Our data platform integrates seamlessly, performs reliably, and delivers at industry-leading speeds. Learn more about Actian, a division of HCLSoftware: www.actian.com