Actian Avalanche

The Revolutionary Gen III Cloud Data Warehouse

A Technical Overview
White Paper

Contents

Introduction:

Since their inception almost 50 years ago, how data warehouses are used has changed drastically. Many enterprises are looking to the data warehouse to maximize their data's potential by hosting data from a wide variety of sources, answering complex queries in real time for business and technical users. Today's data warehouse users expect to do their own data discovery and ad-hoc querying and are no longer satisfied by a set of pre-defined reports and dashboards provided by IT. In following this seismic industry shift, data warehouses have been reimagined from on-premise workhorses, to agile, scalable, flexible solutions in the cloud.

The data management and analytics industry has experienced three notable generations of cloud data warehouses: 1) data-warehouse software that leverages cloud infrastructure; 2) cloud-native, fully managed services; and now, 3) hybrid, multi-cloud fully managed solutions.

Actian Avalanche is a third-generation cloud data warehouse, a fully managed service that provides hybrid capability enabling data to be managed and analyzed simultaneously on-premise and in public clouds. All data can be connected to the broader data ecosystem regardless of location, and organizations may leverage real-time insights that incorporating all of their data assets provides.

From a business perspective, the primary rationale behind moving to the cloud is to reduce cost, particularly with respect to Capital Expenditure. It is also about business agility and speed of innovation. Meeting these goals must be done with the least risk from a delivery and security standpoint. In this white paper, we will describe the underlying architecture of the Actian Avalanche Gen III Cloud Data Warehouse and why it is a superior option for your analytic workloads given today's increasingly varied, demanding, and growing user base and datasets.

Actian Avalanche is designed from the ground up for enterprises that run a wide variety of analytics workloads across multiple cloud and on-premise environments. With Actian Avalanche, CDOs and their data architects and engineers can specify a single hybrid data solution to support standard SQL and ad-hoc analytics workloads with high speed ingestion on the back end and virtually any packaged reporting and visualization tools on the front end, thereby offering a single platform for DBAs to easily maintain and manage services for developers, data scientists, business analysts and virtually any end-user requiring analytics.

Actian Avalanche exploits performance features in today's x86 CPUs that other cloud data warehouses and relational databases do not take advantage of. As a result, Actian Avalanche can process data much faster and deliver analytic workload results quicker than most other relational databases. Much faster

data processing performance opens up opportunities for more iterations during model tuning by data scientists, more "what if" scenario runs for teams of business analysts trying to determine the best business decisions in real-time and so forth. Actian Avalanche delivers support for larger data sets, more users and more complex workloads, as well as the ability to directly query detail data without requiring extensive indexing and materialization of intermediate results.

This paper explains how Actian Avalanche achieves extremely fast performance for analytics workloads and how it can incorporate data from a broad spectrum of data repositories and applications, residing at the edge, in the cloud and existing data centers. But don't just read this paper—experience Actian Avalanche in action. You can start a free trail of Actian Avalanche, load your data and deliver insights in as little as 20 minutes. Or, if you're on the business side or a business analyst and want to understand what to tell your IT folks or need a demo, visit us at www.actian.com/avalanche.

Fully Managed Service

Actian has supported and provided operational services for mission critical database deployments for decades and was one of the first cloud service providers. Actian Avalanche 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 data warehousing service that can be delivered on-premise, in the cloud, or a combination of the two.

Architecture Overview

Figure 1 provides a high-level depiction of the Actian Avalanche architecture which is designed for maximum compatibility with existing workloads and integration with external data sources. Actian's embrace of open standards prevents vendor lock-in, while maximizing compatibility with existing BI and analytics payloads.

Actian and any of its competitor data warehouse vendors will provide you with detailed "how to" documentation to provision and stand up their platforms. In this paper we will focus on a few key differentiators for Avalanche over any of these alternatives.

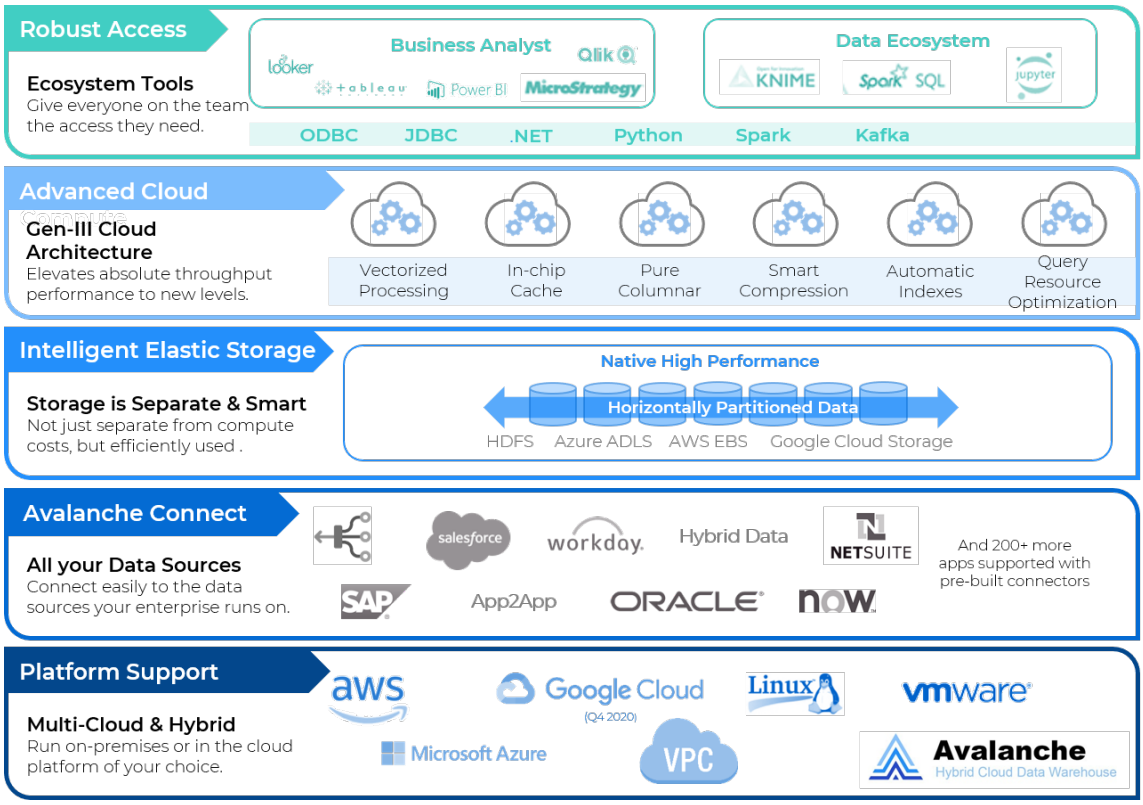

Figure 1. Actian Avalanche Architecture

Avalanche Cloud Compute Acceleration™— Maximizing Cloud Compute Resources

Actian Avalanche was written 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.

Actian Avalanche Cloud Compute Acceleration is unique because it takes advantage of powerful CPU features that most other data warehousing solutions don't, accelerating on-premise and cloud environments alike.

Examples include so-called SIMD[1] 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.

Exploiting Single Instruction, Multiple Data (SIMD)

SIMD enables a single operation to be applied to a set of data at once. Actian Avalanche takes advantage of SIMD instructions by processing vectors of data through the Streaming SIMD Extensions instruction set. Because typical data analysis queries process large volumes of data, using SIMD 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, Actian Avalanche processes vectors of hundreds or thousands of elements at once, which effectively eliminates these overheads. 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, Actian Avalanche avoids using shared RAM as execution memory. Instead, Actian Avalanche uses the private CPU core and CPU caches as execution memory, delivering significantly greater data processing throughput.

---

[1] SIMD stands for Single Instruction, Multiple Data. Traditionally CPUs would process using a SISD model: Single Instruction, Single Data. For more information, see http://en.wikipedia.org/wiki/SIMD.

Avalanche Storage

Actian Avalanche uses cloud storage—and all the resiliency benefits it brings—to keep your data safe. Data volumes are encrypted for security. The data itself is partitioned horizontally to optimize performance. Since the data and compute resources in Actian Avalanche are separated, Actian enables the compute resources to be shut off when not in use, so that you pay only for data storage, and lets you restart them when users come back online.

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 storage model works well for On-Line Transaction Processing (OLTP) systems in which data is stored highly normalized, tables are relatively narrow, queries often retrieve very few rows, and many small transactions can be processed.

Analytics databases are different:

- Tables are often (partially) denormalized, resulting in many more columns per table, not all of which are accessed by most operations.
- Most queries access data from many rows, but result sets are typically small.

Data is added through a controlled rather than ad-hoc process, and often large data sets are added at once or through an ongoing (controlled) stream of data.

As a result of these differences, a row-based storage model typically generates a lot of unnecessary IO 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. In addition to the benefit of data elimination when accessing fewer than all table columns in a query, another significant advantage of column-based storage is better data compression since all of the elements within a column will be of the same data-type.

Automatic Storage Indexes

Actian Avalanche automatically maintains a storage index per column, storing minimum and maximum values 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.

Hybrid Availability

Hybrid capabilities address one critical requirement for many organizations that past generations of cloud data warehouses didn't meet: an on-premise equivalent to the cloud solution that delivers the same managed service and enables the same technologies, skills, and applications for sensitive data management and analytics. This addresses the demand that industries with regulatory compliance requirements, such as financial services, healthcare, and pharma have around their sensitive data.

These enterprises want to leverage the same technologies for their analytics needs to write applications that seamlessly join on-premise and cloud-resident data. Actian Avalanche 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 seamlessly blended to provide 360-degree customer insights.

Real-time Update Capability

A big challenge with most column-based databases is incremental small inserts, updates, or deletes (as opposed to large bulk data load operations). Actian Avalanche meets this challenge with high-performance in-memory Positional Delta Trees (PDTs). Actian Avalanche 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. 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.

Actian Avalanche implements a fully ACID[2]-compliant transactional database with multi-version read consistency. Any new transaction will see all previously committed transactions, both small incremental transactions and large bulk data loads. Changes are always written persistently to a transaction log before a commit completes to ensure full recoverability.

Data Compression

Actian Avalanche compresses data on a column-by-column, page-by-page

[2] ACID stands for Atomicity, Consistency, Isolation, Durability—a set of properties that guarantees database transactions are processed reliably. For more information, visit http://en.wikipedia.org/wiki/ACID.

basis using any of the following algorithms or a combination of them:

- Run-Length Encoding (RLE)[3]: 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 data type. 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. 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.
- LZ4: This method detects and encodes common fragments of different string values. It is particularly efficient for medium and long strings.

The algorithms Actian Avalanche uses to compress data were selected for their speed of decompression over a maximum compression ratio. The compression ratio you can achieve with Actian Avalanche is highly data-dependent: 4—6x compression ratios are common for real-world data but both lower and higher compression ratios have been observed.

Parallel Execution

Actian Avalanche implements a flexible adaptive parallel execution algorithm. Actian Avalanche can execute statements in parallel using any number of CPU cores and will intelligently balance concurrency and query parallelism.

Actian Avalanche Data Ingestion

Actian Avalanche takes analytic data processing to a new level. With it, you can now achieve amazing performance with a simple, ANSI-compliant relational database—something previously achievable only with expensive on-premise

[3] See http://en.wikipedia.org/wiki/Run-length_encoding.

data warehouses or following lots of careful design and tuning using complex features.

Use Actian Avalanche if you are looking for a relational database, that supports ANSI SQL and industry-standard JDBC/ODBC interfaces. Actian Avalanche delivers extremely fast performance, is easy to use, and is very cost-effective. Actian Avalanche delivers performance faster than popular in-memory type databases without even having to load all data in memory and without the hard limit of available memory. Figure 2 shows a number of areas where you should consider Actian Avalanche.

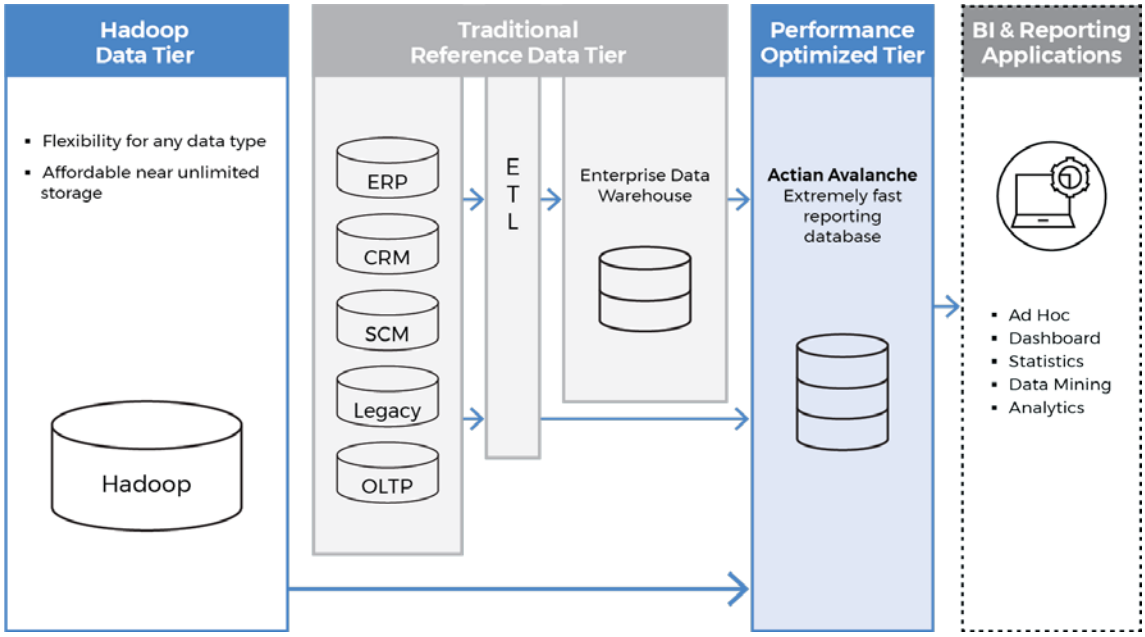

Figure 2. Cooperative Analytics with Actian Avalanche

Avalanche FlexPath™ Connectivity

Avalanche with its built-in comprehensive data access supports including SPARK and Kafka open source functionality. Pre-built connectors can provide a broad range of high-performance access to today's most popular applications and data sources (add how it works, how it is different, how it benefits).

Avalanche Hybrid Data Federated Query Support

how Avalanche and Vector can seamlessly execute queries that pull from one or more Actian-based analytics databases on-premise or in the cloud without an ETL tool or forcing data movement across sources to support hybrid data-based analytics (add how it works, how it is different, how it benefits).

Hadoop Data Lake Support

Just as legacy enterprise data warehouses are slow to change or retire, given the number of applications and users in any given enterprise leveraging them. Hadoop also will be with us for years to come because of its inherent low-cost and high-reliability storage with HDFS, the Hadoop storage layer (YARN, the Hadoop management layer is also seen as a long-term tool). However, the other tools in the Hadoop family for processing the data, MapReduce, Hive, Pig, Mahout, etc., have far superior analogues both on-premises as well as in the cloud. Actian Avalanche acts as a superior data processing platform in the cloud or on-premise through either external tables or large-scale extraction.

Actian Avalanche's external table capability enables you to register files, such as Parquet or ORC files sitting on HDFS, as tables within the database and to join the data from those files with native Actian Avalanche data. External tables capability provides push-down to minimize the data being brought back from the remote source. The use of a "CREATE TABLE AS SELECT…" statement is a simple way to ingest the entire contents or a subset of the external file.

Alternatively, Actian Avalanche supports using Spark as a high-speed transfer mechanism between HDFS-based data lakes and Avalanche instances, thereby enabling large-scale streaming data extraction from Hadoop data lakes for raw data and insertion of data analysis results.

By using these options, Avalanche customers can leverage existing investments in their data lakes and continue using them with other projects more suited to Hadoop while off-loading high-speed, interactive use cases such as market basket analysis, AI/ML algorithm tuning, and other end-user applications for business analysts and data scientists.

High Speed Data Ingestion Support

Actian Avalanche 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; others prefer to use Spark for high-speed ingestion of large volumes of data.

You can perform traditional ETL operations including data transformation, data quality, and data cleansing in parallel, eliminating traditional performance bottlenecks in your data acquisition processes. Moreover, you can use Actian 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.

Actian Avalanche Security

There was a time when public cloud offerings were perceived as less secure and more vulnerable than traditional on-premise security. The accepted best practices for enterprise security were well-defined firewalls and intrusion detection and prevention systems, paired with access control points, user behavior, and authorization, authentication, and access management security.

However, while the number and variety of threats across the board against any IT system connected to a network continue to rise, distributed bot attacks explicitly targeting Cloud Vendors has surpassed others. It's therefore counterintuitive that most major breaches attributed to attacks are to on-premise and poorly designed private cloud deployments. Over time, the cumulation of expected versus actual successful attacks has meant that cloud security for the major cloud players like Amazon, Microsoft, and Google are gaining reputations as more secure based on track records.

Much of the success cloud service providers have in defending against these distributed attacks can be attributed to continuous diagnostics and monitoring combined with mediation. Attacks based on methodical and pre-defined scenarios have led to willingness to place data in the cloud. In this section, we will delve further into the Avalanche enterprise security framework.

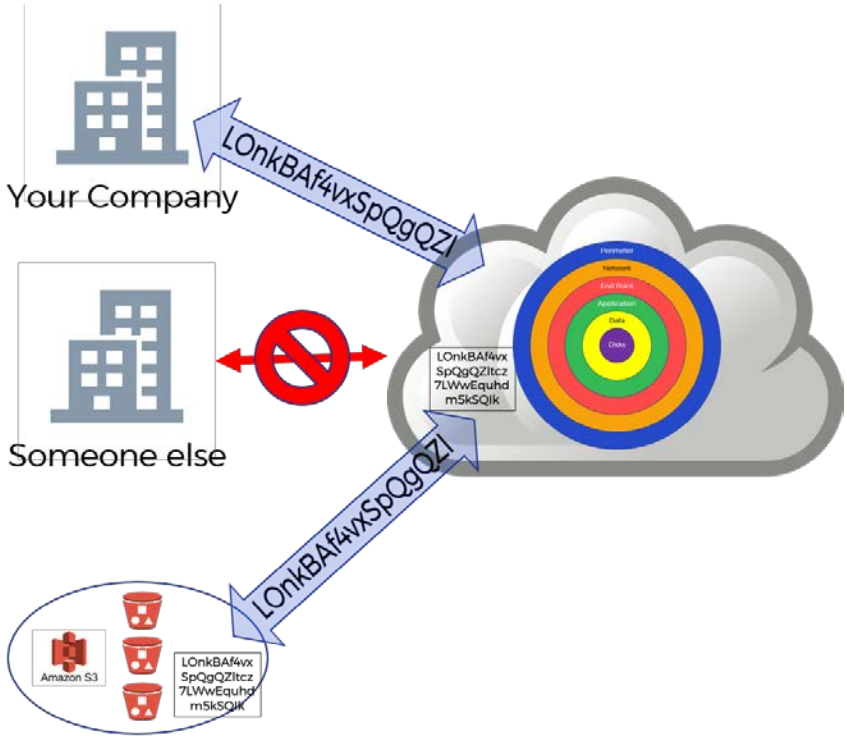

Figure 3. Avalanche Cloud Security Framework, Key Features

Actian Avalanche Cloud Security Framework

Modern cloud security posture is like an onion, with layer upon layer of security in each cloud service. Each layer is monitored and alerted so any penetration is addressed immediately. As mentioned previously, the three major cloud service providers deliver a solid layer of security for the underlying infrastructure and supporting services to PaaS and SaaS offerings residing on their platforms. Actian Avalanche leverages these services.

Additionally, the Actian Cloud Security Framework includes single and multi-tenant data management features; white-listing and non-public facing IP/ports; user-, group- and role-based access control; data encryption including column-level data at rest encryption on top of the encrypted file-system, data masking; and dynamic and ongoing security auditing and security alarms.

Access Control and Client Data Management

Each Actian Avalanche customer's environment is provisioned and managed by Actian Corporation. During provisioning, Avalanche assigns each customer's environment its own Virtual Private Cloud to remove any risk of data leakage across customers. Avalanche leverages the VPC facilities from AWS to enable gated access to each client's environment through a whitelist of IP addresses. To further ensure security and complete isolation between clients, whitelisted IP addresses should be static—either directly or indirectly through the VPC where dynamic IP addresses are employed. All data, whether in-motion or at rest, is encrypted by default.

After the environments are provisioned, clients can use standard LDAP or Active Directory authentication and authorization vehicles to establish user accounts against the whitelisted IP addresses. Standard secure transport layer mechanisms, for example SSL, are then used to bind the LDAP/AD connections. Avalanche leverages AWS Key Management Services to create and manage keys and control the use of encryption by Avalanche services such as JDBC and ODBC or Spark ingestion.

Data Security

Actian Avalanche incorporates the underlying cloud provider's data security. For example, with AWS, Avalanche leverages AES 256-bit encryption for data at rest and FIPS 140-2 for key encryption as data payloads are transferred to and from S3, ADLS, Google Cloud Storage and the Avalanche instance or to and from external customer repositories. Storage within the Avalanche instance is further secured through file system encryption with block-level corruption detection. If Avalanche detects a corrupted block, it will replace the block.

Actian Avalanche Analytics Workload Support

Actian Avalanche is based on an architecture that satisfies the requirements of your entire virtual team, deftly meeting the needs of a wide range of roles and associated skillsets from data architects and engineers to data scientists, business analysts, and power users across various lines of business. Actian Avalanche can quickly support on top of or connect to a wide range of programming languages. It provides full SQL compliance and SQL tools support, Python, Jupyter, R-Lib, and TensorFlow, as well as virtually any visualization and reporting tool—from Microsoft PowerBI to Tableau.

SQL Support

Actian Avalanche supports the SQL 2016 standard including analytics capabilities such as CUBE, ROLLUP, GROUPING SETS, and analytic windowing functions.

The primary benefit of open standards is to prevent vendor lock-in. Actian Avalanche customers may immediately leverage this by pointing their existing workloads at Actian Avalanche. It typically works—only a lot faster, without requiring any changes.

Universal Development and Deployment Model

Business analysts, power users, even data scientists and developers need faster performance and the ability to rapidly stand up new projects or rapidly extend existing ones to larger or new data sources. They want a platform that is not seen as burdensome or costly in terms of schedule and resource consumption.

Actian Avalanche prevents these stakeholders from asking their DBAs and IT operations to deliver the impossible on aging legacy systems. It also provides better price-performance than any other cloud data warehouse alternative.

Conclusion

Actian Avalanche 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.

Actian Avalanche 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 Avalanche performs.

Getting started with Actian Avalanche is:

1. Create your Avalanche cluster.
2. Load your data.
3. Start querying.

Your production cluster could be up and serving your business needs in as little as 20 minutes.

Actian Avalanche 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.

Actian Avalanche provides a single platform for business analysts, developers, data scientists, and data engineers. Avalanche enables an enterprise 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 Avalanche, your enterprise will foster faster innovation. Now you can shorten AI and machine-learning life cycles with parallelized data loads. Because of the performance it delivers, no sampling is required. And Avalanche's open architecture integrates with R, Python, Spark, and dozens of other tools.

There is nothing as compelling as seeing your own queries running against your own data. Prepare to be amazed. Get started with Actian Avalanche today, for free.

About Actian – Activate your Data™

Actian, the hybrid data management, analytics and integration company, delivers data as a competitive advantage to thousands of customers worldwide. Through the deployment of innovative hybrid data technologies and solutions, Actian ensures that business critical systems can transact and integrate at their very best—on premise, in the cloud or both. Thousands of forward-thinking organizations around the globe trust Actian to help them solve the toughest data challenges to transform how they run their businesses, today and in the future. For more, visithttps://www.actian.com.

2300 Geng Rd, Suite 150, Palo Alto, CA 94303 +1 888 446 4737 [Toll Free] | +1 650 587 5500 [Tel]