acceldata

www.acceldata.io

xLake Reasoning Engine Performance Benchmark Report

Profiling and Validation with TPC-DS Dataset on Google Cloud Platform (GCP)

AICPA SOC

HIPAA COMPLIANT

ISO 27001

Acceldata xLake Reasoning Engine Performance Team

First Edition - Oct 2025

Abstract

Understanding and assuring data quality is critical for organizations that make data-informed decisions. Data quality means both completeness and accuracy, and is increasingly crucial for aligning with industry standards and regulatory requirements. As enterprises rely ever more on data for operations, strategy, and innovation, trust in data is paramount.

Reliable healthcare data improves patient outcomes; financial organizations depend on it for predictions and risk assessments; and e-commerce companies rely on trustworthy data for inventory management and customer experience.

Ensuring quality at scale on cloud platforms such as GCP presents unique challenges. Processes that work for small datasets may falter when scaled, jeopardizing data quality and integrity.

We used the TPC-DS dataset in our study to evaluate a data dependability product. TPC-DS, developed by a group of industry specialists, serves as a model for decision-making systems. Such systems, known for their stringent data processing needs, sift through massive amounts of data in order to extract useful insights. They are adaptable, able to handle a wide range of data jobs, from simple searches to in-depth data mining.

We present a detailed look at our benchmarking approach in this paper, addressing the specific issues posed by the TPC-DS dataset and the strategies we used to overcome them. We intend to give a complete, technical overview of data reliability evaluation while emphasizing validation's critical significance as a cornerstone in today's data-centric landscape.

Contents

Introduction

Background

Modern organizations rely deeply on data for strategic decisions, operational pivots, and forecasting. This dependency emphasizes the critical need for data quality—accurate, complete, and trustworthy information infrastructure. Without this, insights can be misleading, resulting in costly errors and risks.

Achieving quality is increasingly challenging at scale. Validating vast, complex datasets efficiently is a significant technical hurdle. Our benchmarking project leverages the industry-standard TPC-DS dataset, rigorously assessing our platform's quality and providing a reference point for customers.

Objective

Given the challenges and complexities surrounding data validation and the consequential implications for data reliability, the Acceldata Engine Performance team set out to create a robust, industry-acknowledged benchmark to evaluate our platform's data reliability product. This motivated us to leverage the TPC-DS dataset, established by a third-party committee comprising industry specialists. TPC-DS has emerged as the gold standard for evaluating decision-support solutions.

Performing this benchmark offers dual advantages— it allows us to subject our product to a rigorous test environment that mirrors real-world complexities and provides our customer community with an understandable reference point for performance. Through this study, we aim to not only assess our product's proficiency in ensuring data reliability but also to deepen the collective understanding of data validation's nuances at scale.

Importance of Trusted Data in AI Initiatives

Reliable data is critical for making corporate choices, improving operational efficiency, ensuring regulatory compliance, and building user confidence. Inconsistent or erroneous data can result in inefficiency, squandered resources, and legal ramifications. It is critical for stakeholders to ensure data integrity and confidence.

Reliable Analysis: Business decisions are frequently based on insights derived from data. If the underlying data is erroneous, the resulting decisions could be flawed, leading to potentially costly mistakes.

Regulatory and Compliance: Many industries are governed by strict data-related regulations. Ensuring data integrity and reliability helps organizations stay compliant and avoid potential legal repercussions.

Operational Efficiency: Clean, reliable data streamlines operations. Inconsistent or incorrect data can lead to process inefficiencies, increased manual intervention, and wasted resources.

User Trust: For end-users, analysts, or stakeholders, trust in the data is paramount. Reliability rules ensure that data is consistent and trustworthy, bolstering user confidence.

Challenges at Enterprise Scale

Enterprises often manage vast data from diverse sources, making data reliability challenging. Dynamic data sources and performance overheads add to the challenge, making it difficult to balance performance with reliability.

Volume and Complexity: Enterprises often deal with vast amounts of data from varied sources. Ensuring reliability across such large datasets is resource-intensive and can be challenging to manage.

Performance Overhead: Implementing strict data reliability checks can add performance overheads to the data ingestion and processing pipelines. Balancing performance with reliability becomes a challenge.

Dynamic Data Sources: At a large scale, data sources might frequently change, evolve, or get updated. Keeping up with these changes while maintaining data reliability can be demanding.

Approach

Platform Selection

We used the Google Cloud Platform (GCP) for our benchmarking, with BigQuery serving as the primary datastore GCP delivers a dependable and scalable infrastructure, while BigQuery, which is known for its quick SQL analytics across massive datasets, provides an ideal environment for precise testing and evaluation.

Dataset Generation

We employ a synthetically generated TPC-DS dataset with nearly 6.3 billion rows spread between fact and dimension tables for this benchmark, as described in the table below.

Table	Table Type	Rows (Millions)	Columns
store_sales	Fact	2879 M	22
catalog_sales	Fact	1439 M	33
inventory	Fact	783 M	3
web_sales	Fact	719 M	33
store_returns	Fact	288 M	19
catalog_returns	Fact	144 M	26
web_returns	Fact	72 M	23
customer	Dimension	12 M	18
customer_address	Dimension	6 M	13
customer_demographics	Dimension	1.9 M	9
item	Dimension	0.3 M	22

Performance Metrics

Our evaluation is anchored in four primary performance metrics:

Data Profiling: Data Profiling is the process of studying and analyzing datasets in order to collect descriptive statistics and understand their structure, content, and quality. It examines data properties, relationships, abnormalities, and potential contradictions to verify that it is ready for further processing or analysis.

Speed: The rate at which data is validated is measured as speed. Given the contemporary corporate landscape's rapid decision-making demands, speedy validation is critical.

Data Validation Accuracy: This metric assesses the tool's ability to find and correct data discrepancies, ensuring that the data remains true to its source and intended usage.

Reliability: Reliability indicates the constancy of the tool's performance regardless of the amount or complexity of the dataset.

- The metrics employed for this evaluation are relevant to the needs of today's data driven enterprises.
- Accuracy in data validation ensures that businesses can derive genuine insights from their data.
- Reliability assures enterprises of the tool's dependable performance for continuous operations.
- Profiling makes sure that enterprises understand the shape of their data and can make informed decisions on down stream processing.
- Speed invalidation enables more data to be pushed through the validation process.

Methodology

Data profiling and reliability standards applied to fact and dimension tables, such as those in the TPC-DS (Transaction Processing Performance Council - Decision Support) benchmark dataset, provide a structured method for ensuring data accuracy, consistency, and reliability. These rules include, among other things, referential integrity, domain integrity, and uniqueness constraints.

Ruleset

The rule sets shared earlier for the TPC-DS dataset are quintessential to ensuring data reliability. These rules are a combination of accuracy, consistency, and clarity, which aligns with the overall objective of achieving dependable and interpretable data

Referential Integrity: The use of ‘Lookup’ rules, where columns in the store_sales fact table reference primary keys in corresponding dimension tables, is vital for ensuring referential integrity. This guarantees that relationships between tables are maintained, ensuring that the data in the fact table corresponds to valid, existing data in the dimension tables.

Domain Integrity: By imposing SQL rules such as “non-negative” constraints on numerical columns, the dataset upholds domain integrity. These rules ensure that the values in these fields adhere to realistic and expected bounds, eliminating potential anomalies that could skew analyses.

Data integrity: Rules like “NULL Uniqueness” or specific uniqueness constraints on identifiers ensure that each entry or transaction is distinct. This is crucial for preventing data duplication, which could lead to inflated or misleading metrics in subsequent analyses.

Here is an example of a set of rules applied to the store_sales tables. For an exhaustive set of rules, please refer to the following appendix.

Column	Rule in English
ss_sold_date_sk	Must match an entry in the `date_dim` table’s `d_date_sk` column.
ss_sold_time_sk	Must match an entry in the `time_dim` table’s `t_time_sk` column.
ss_item_sk (1)	Cannot be NULL, must be unique, and cannot have duplicates.
ss_customer_sk	Must match an entry in the `customer` table’s `c_customer_sk` column.
ss_cdemo_sk	Must match an entry in the `customer_demographics` table’s `cd_ demo_sk` column.
ss_hdemo_sk	Must match an entry in the `household_demographics` table’s `hd_ demo_sk` column.
ss_addr_sk	Cannot have any rule and can be NULL.

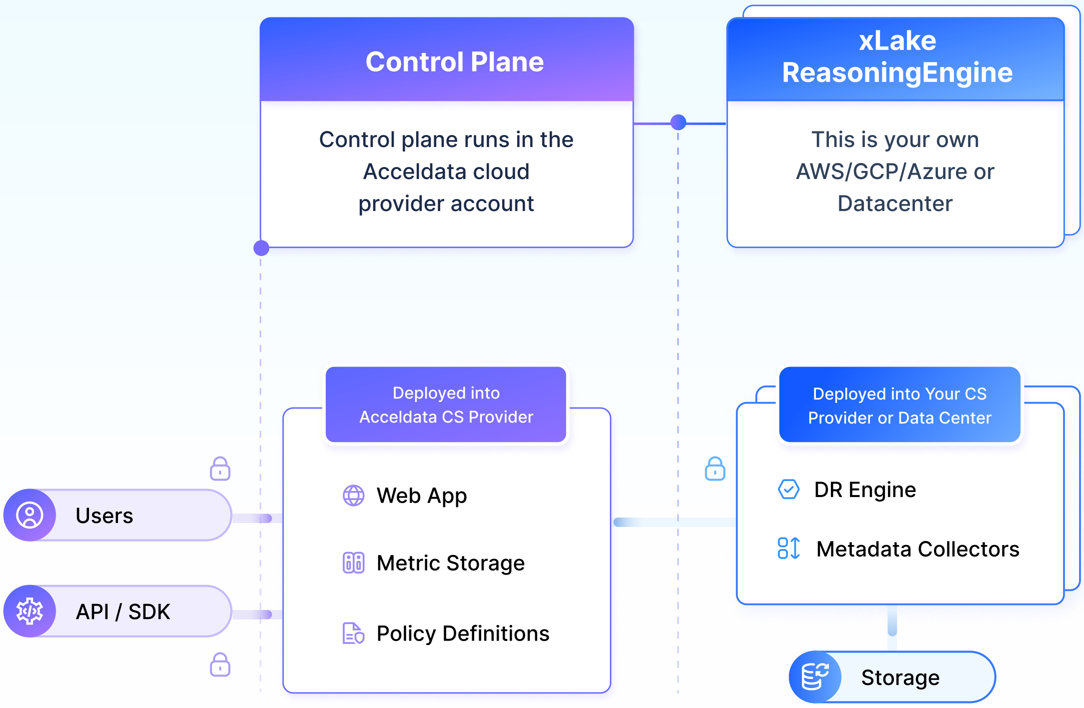
Column	Rule in English
ss_store_sk	Must match an entry in the \`store\` table’s \`s_store_sk\` column.
ss_promo_sk	Must match an entry in the \`promo\` table’s \`p_promo_sk\` column.
ss_ticket_number (2)	Cannot be NULL and must be unique.
ss_quantity	If provided, the value must be non-negative.
ss_wholesale_cost	If provided, the value must be non-negative.
ss_list_price	If provided, the value must be non-negative.
ss_sales_price	If provided, the value must be non-negative.
ss_ext_discount_amt	If provided, the value must be non-negative.
ss_ext_sales_price	If provided, the value must be non-negative.
ss_ext_wholesale_cost	If provided, the value must be non-negative.
ss_ext_list_price	If provided, the value must be non-negative.
ss_ext_tax	If provided, the value must be non-negative.
ss_coupon_amt	If provided, the value must be non-negative.
ss_net_paid	If provided, the value must be non-negative.
ss_net_paid_inc_tax	If provided, the value must be non-negative.
ss_net_profit	If provided, the value must be non-negative.

Execution Environment

Acceldata manages the ADOC control plane cloud service, which is secure, multi-tenanted, and SOC2-certified. Authentication, user account management, job orchestration, and system-wide monitoring are all handled by the ADOC central management hub. The control plane collects and saves metrics, allowing users to create and manage data reliability policies. It manages user accounts in order to limit system access to only authorized users. The control plane constantly checks the system for anomalies.

xLake Reasoning Engine is a high-performance compute engine with a single tenant for high-volume data processing. It performs all data processing and scales to petabytes using Apache Spark and Kubernetes. The combination of Spark and Kubernetes improves scalability, resource efficiency, and isolation. Data validation efforts in the data plane are powered by a 12-node GKE cluster.

Acceldata Architecture


Control Plane
Control plane runs in the Acceldata cloud provider account
xLake ReasoningEngine
This is your own AWS/GCP/Azure or Datacenter
Deployed into Acceldata CS Provider
Web App
Metric Storage
Policy Definitions
Users
API / SDK
Deployed into Your CS Provider or Data Center
DR Engine
Metadata Collectors
Storage

Control Plane

The ADOC control plane is a secure, multi-tenanted, SOC2-certified cloud service operated and managed by Acceldata. The control plane is the central management hub of the ADOC platform, taking care of essential functions such as authentication, user account management, task orchestration, and system-wide monitoring.

The control plane serves as a unified system layer that abstracts complexities and provides a simplified interface to users, ensuring they can focus their attention on crucial data processing and observability tasks.

Key Functions of the Control Plane:

User Interface and APIs: Serving as the primary interaction point for users, the control plane provides a user-friendly interface and comprehensive APIs. This functionality allows users to effectively interact with the system, initiate jobs, monitor task statuses, and more, without delving into the system's underlying complexities.

Metrics collection and storage: The control plane gathers metrics from the various connectors that ADOC provides and manages the storage and processing of these metrics. It can run ML and AI algorithms on top of this data to provide insights into your compute and storage infrastructure.

User Account Management: The control plane manages all user accounts in the system, including account creation, user roles and permissions, and account termination. This comprehensive user account management ensures that only authorized personnel can access the system and its various features.

Authentication and Security: Being a multi-tenanted system, the control plane places heavy emphasis on stringent security measures. As a SOC2 certified entity, it upholds high standards of security, availability, processing integrity, confidentiality, and privacy. It governs user access control, audit trails, and logs to ensure system-wide integrity and security.

Data Reliability Policy Authoring and Management: Users Author and manage their data reliability policies on the control plane UI, which provides a low-code and code-based interface to create and manage data reliability policy lifecycle.

Data Reliability Policy Authoring and Management: The control plane consistently monitors the system for potential issues or anomalies. It generates alerts for failures or unusual activities, facilitating timely detection and resolution of problems

Through these critical functions, the control plane simplifies the usage of the software, manages essential system-wide administrative tasks, and provides a simplified, secure interaction layer for users.

xLake Reasoning Engine

The xLake Reasoning Engine, built on the foundation of the Acceldata Data Observability Cloud (ADOC), extends the platform with an intelligence layer deployed in the customer's environment (AWS, GCP, Azure, or on-premises datacenters). While the Control Plane—managed by Acceldata—hosts the web application, metric storage, and policy definitions, xLake operates alongside customer data systems to execute reliability checks, collect metadata, and perform large-scale distributed processing. This separation ensures secure, high-performance workloads close to the data while leveraging the Control Plane for orchestration and governance. By unifying observability signals with agentic AI reasoning, xLake transforms raw telemetry into actionable insights, accelerating root-cause analysis, anomaly detection, and cost optimization.

Scalability: Kubernetes allows Spark to scale in and out based on workload. It can quickly spin up pods to handle the increasing load and just as quickly spin them down when they're no longer needed. This elastic scaling can be particularly cost-effective in cloud environments where you pay for what you use.

Isolation and Security: Kubernetes provides isolation between Spark jobs running in different pods. This isolation can improve security and prevent one job from interfering with another.

Resource Efficiency: With Kubernetes, you can achieve better resource efficiency. You can run multiple Spark jobs on the same Kubernetes cluster and share resources among them, which is often more efficient than running each job on its own dedicated cluster.

The ADOC data plane provides a robust and scalable solution for high-volume data processing and metadata management and can support a range of diverse data reliability requirements.

Cluster Sizing

The data plane environment was provisioned with a 12-node GKE (Google Kubernetes Engine) cluster. This infrastructure offers the foundational resources we require to run our Data validation jobs.

Executor Configuration

Executor Cores:

Each executor we've configured is allocated 5 cores.

Executor Memory:

24GB of memory for each executor.

Overhead Memory:

We've set aside an additional 5GB for overhead, which covers off-heap memory allocations, both in the executor and driver. This includes necessities like native libraries and metadata of RDDs.

Data and Processing:

The data is divided into 200 partitions.

A distribution of 2 executors across each of 10 nodes, summing up to 20 active executors for our job.

With our given executor resources, our concurrent processing capability stands at 100. This is derived from the collective cores across all executors: 20 executors x 5 cores each = 100 cores in total. This denotes that at any specific moment, we can run 100 tasks concurrently, where each task processes one partition.

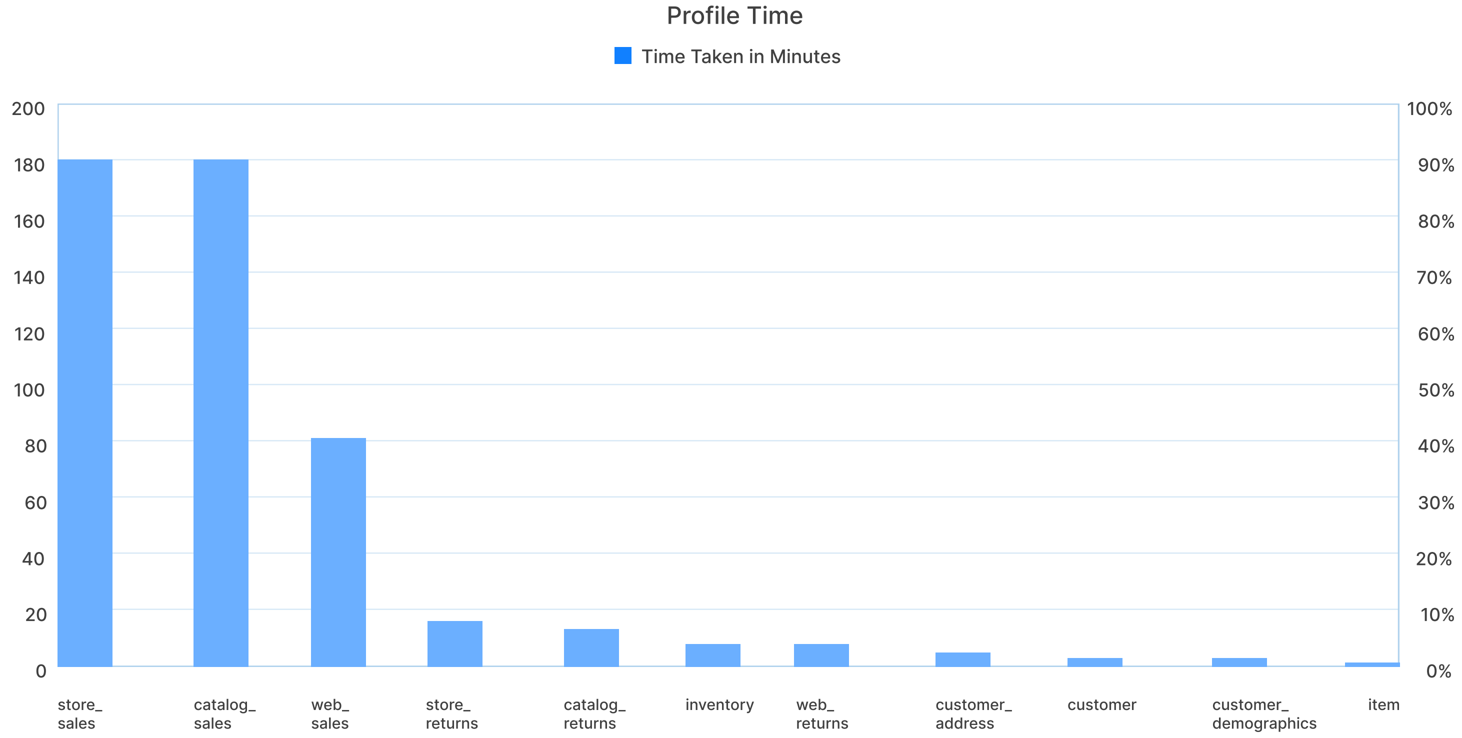
Concurrency:

With our 100 cores available across all executors and our data segmented into 200 partitions, our setup operates on a 2:1 ratio of partitions to cores. This suggests that during the processing of our job, we'd ideally need two cycles of processing to cover all the data.

Results

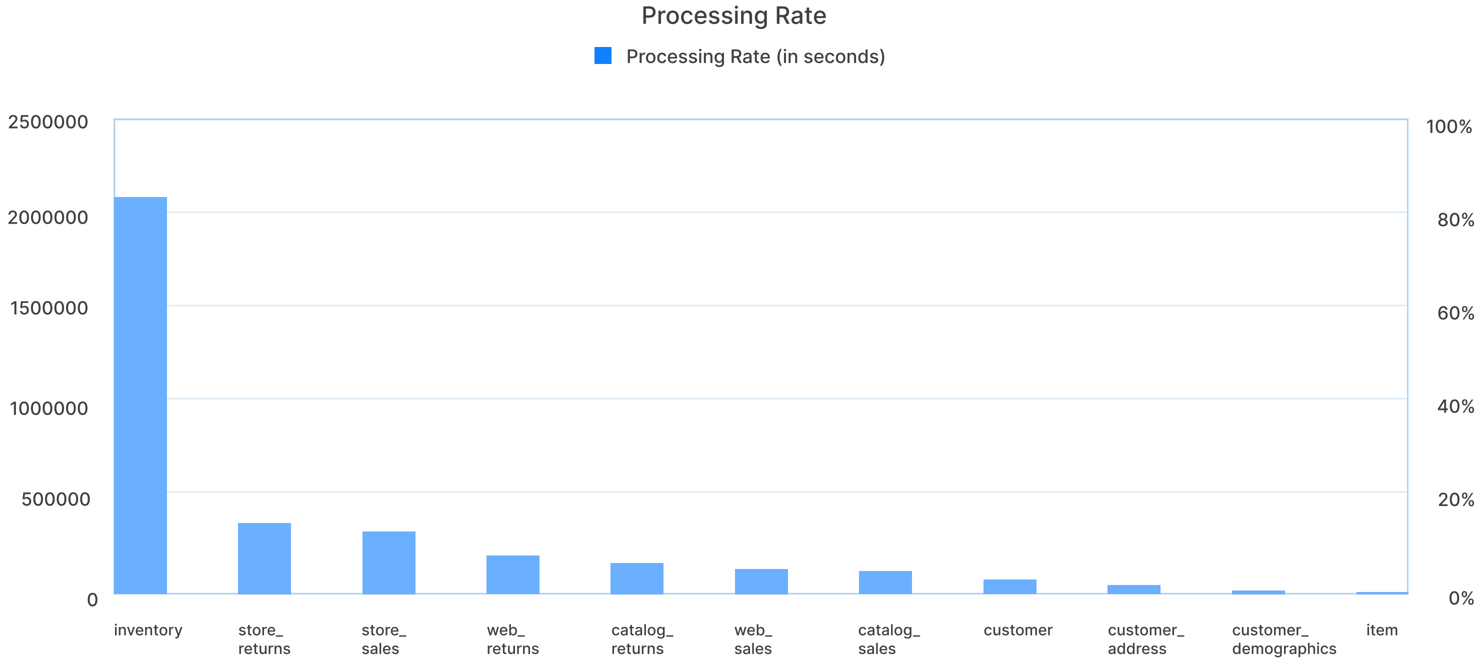
Data Profiling

Table	Data Set Type	Rows	Time Taken (in mins)
store_sales	1 TB	2879966589	180
catalog_sales	1 TB	1439976202	180
inventory	1 TB	783000000	6
web_sales	1 TB	719959864	81
store_returns	1 TB	288009578	16
catalog_returns	1 TB	144004725	13
web_returns	1 TB	72002305	6
customer	1 TB	12000000	2
customer_address	1 TB	6000000	3
customer_demographics	1 TB	1920800	2
item	1 TB	300000	1



Data Profiling

Table	Data Set Type	Rows	Processing Rate(in secs)
store_sales	1 TB	2879966589	266663.5731
catalog_sales	1 TB	1439976202	133331.1298
inventory	1 TB	783000000	2175000
web_sales	1 TB	719959864	148139.8897
store_returns	1 TB	288009578	300009.9771
catalog_returns	1 TB	144004725	184621.4423
web_returns	1 TB	72002305	200006.4028
customer	1 TB	12000000	100000
customer_address	1 TB	6000000	33333.33333
customer_demographics	1 TB	1920800	16006.66667
item	1 TB	300000	5000



Notes on the performance benchmark results for

Dataset Profiling:

Store Sales vs. Catalog Sales:

Both the store_sales and catalog_sales tables in the TPC-DS dataset recorded a profiling duration of 180 minutes. store_sales processed roughly twice the volume of rows compared to catalog_sales. This confirms that the processing efficiency for in-store sales data (store_sales) surpasses that of catalog-based sales data (catalog_sales), given its capacity to handle larger volumes in the same time frame.

Efficiency:

The inventory table in the TPC-DS dataset showcased superior efficiency by processing 783 million rows in only 6 minutes. In contrast, the web_sales table, which captures online transactions, required 81 minutes for a slightly fewer number of rows, confirming the more intricate nature of online sales data.

Dimension Table Profiling:

Dimension tables in the TPC-DS schema, specifically customer, customer_address, customer_demographics, and item, were profiled swiftly, ranging from 1 to 3 minutes. This underlines the optimized nature of these tables, with streamlined data types and structures ensuring efficient profiling.

Returns Data Profiling:

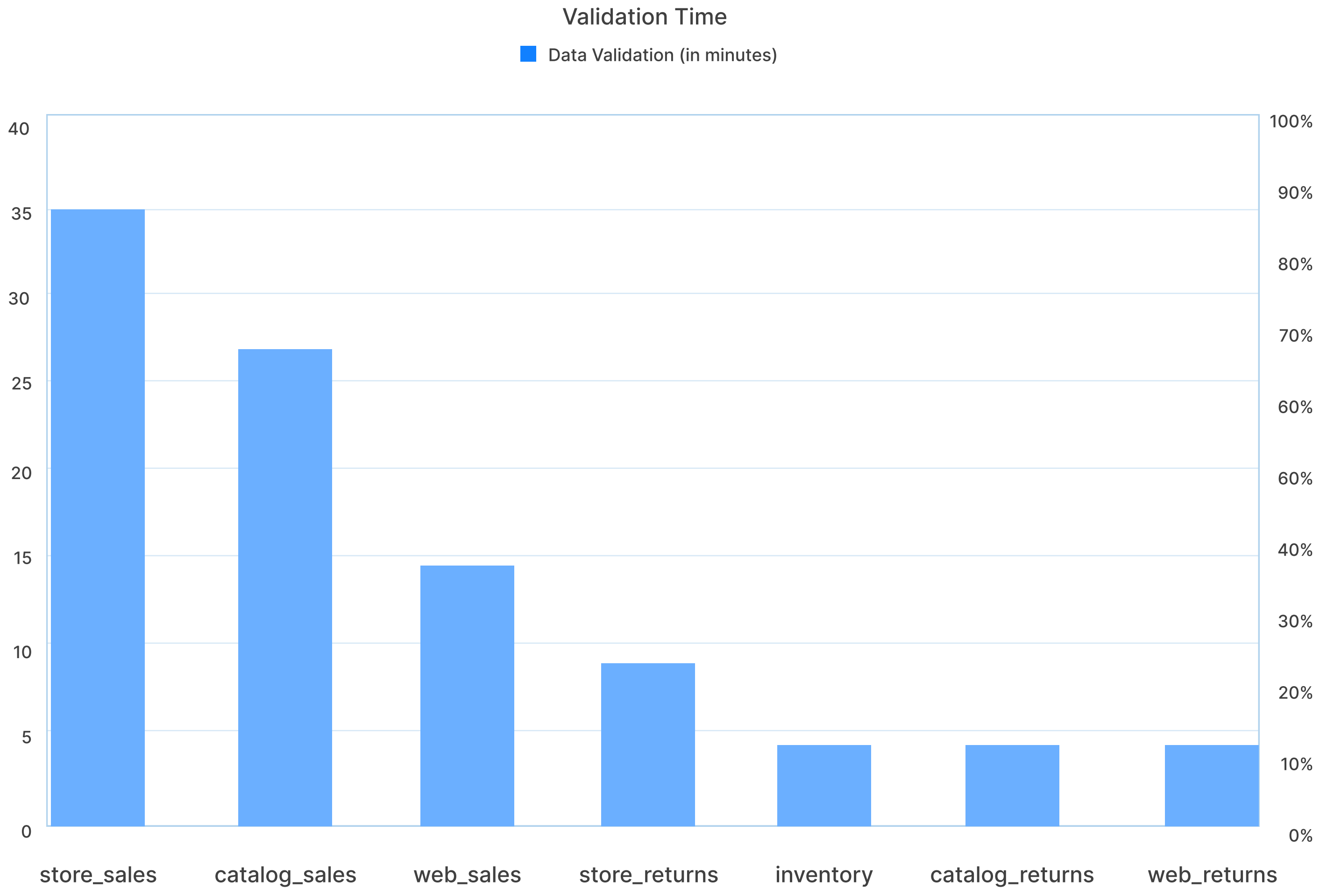
Profiling results for the store_returns and catalog_returns tables in the TPC-DS dataset exhibited minor differences in processing times, despite a four- fold disparity in row counts. This attests to the consistent data structure and datatype distributions in these return records across different sales channels.

Item Data:

The item table in TPC-DS, responsible for product details, registered an exceptional profiling performance, completing the task in a mere 1 minute for 300,000 rows. This indicates the table's simple structure and data uniformity.

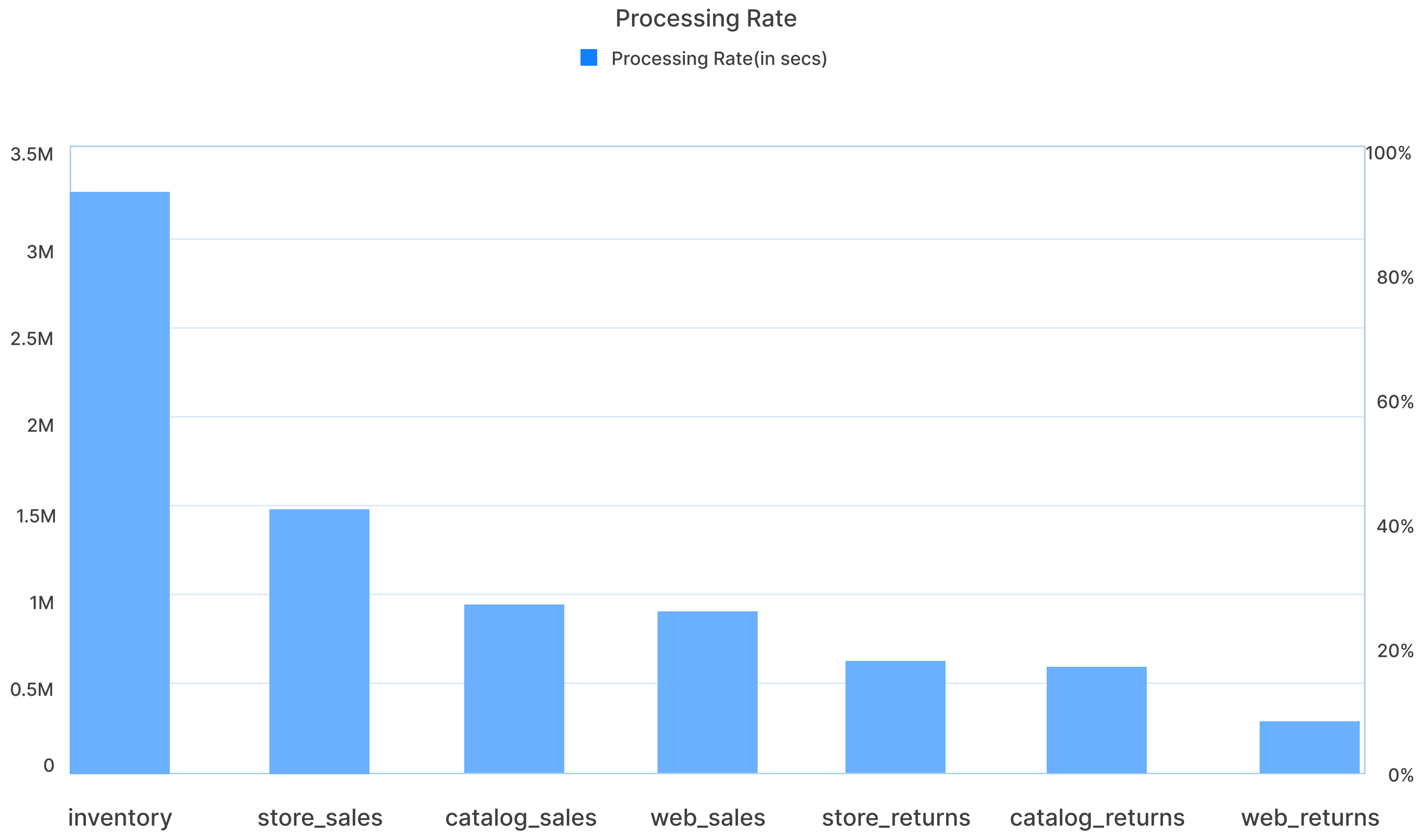
Data Validation

Table	Data Set Type	Rows	Time Taken (in mins)
store_sales	1 TB	2879966589	35
catalog_sales	1 TB	1439976202	27
inventory	1 TB	783000000	4
web_sales	1 TB	719959864	14
store_returns	1 TB	288009578	8
catalog_returns	1 TB	144004725	4
web_returns	1 TB	72002305	4



Data Validation

Table	Data Set Type	Rows	Processing Rate(in secs)
store_sales	1 TB	2879966589	1371412.661
catalog_sales	1 TB	1439976202	888874.1988
inventory	1 TB	783000000	3262500
web_sales	1 TB	719959864	857095.0762
store_returns	1 TB	288009578	600019.9542
catalog_returns	1 TB	144004725	600019.6875
web_returns	1 TB	72002305	300009.6042



Notes on the performance benchmark results for

Dataset Validations:

Store Sales:

With approximately 2.88 billion rows, the data quality checks for the store_sales table were completed in 35 minutes. This translates to a processing rate of approximately 82.3 million rows per minute.

Catalog Sales:

The catalog_sales table processed its roughly 1.44 billion rows in 27 minutes, yielding a processing rate of about 53.3 million rows per minute.

Inventory:

This table stands out for its efficiency. For 783 million rows, the data quality checks were swiftly completed in 4 minutes. This denotes a remarkable processing rate of approximately 195.75 million rows per minute.

Web Sales:

Processing approximately 720 million rows, the web_sales table concluded its data quality checks in 14 minutes. This gives a processing rate of about 51.4 million rows per minute.

Store Returns:

The store_returns table demonstrated efficient processing for its 288 million rows, finishing in 8 minutes. This calculates to a rate of 36 million rows per minute.

Catalog Returns:

With 144 million rows, the catalog_returns table mirrored the efficiency of store_returns, completing in 4 minutes, at a rate of 36 million rows per minute.

Web Returns:

The web_returns table processed its 72 million rows in 4 minutes. This consistent processing speed matches that of catalog returns, with both at 18 million rows per minute.

Summary

In this benchmarking exercise, we delved into the capabilities of ADOC's data plane, which is grounded in its Cloud-native Spark-based architecture. The objective was to measure its prowess in data profiling and reliability, specifically on the extensive TPC-DS dataset with over 6 billion rows

The platform's proficiency was evident as it profiled a whopping 6 billion rows, reaching a remarkable peak speed of 2 million rows per second.

Data validation workloads stood out, achieving a rate of approximately 195.75 million rows every minute, underlining its adeptness at efficiently processing expansive datasets.

Furthermore, when confronted with the massive datasets of primary fact tables like store_sales and catalog_sales, the system upheld its stellar performance. These tables clocked in at 82.3 million and 53.3 million rows per minute, respectively.

The consistent high performance across diverse table sizes and complexities underscores the robust and adaptable nature of ADOC's data plane.

Its versatility and programmability make it invaluable for enterprises that often have to deal with data validation over massive datasets. It enables organizations to swiftly and accurately gain insights from vast data repositories.

Learn more

For more information, see the full documentation on our documentation pages. ADOC is an enterprise-ready cloud platform that is built using a strong security posture for organisations small and large, and across all industries.

We're happy to discuss your specific needs in more detail - info@acceldata.io

About Acceldata

Acceldata, founded in 2018, provides the industry's first unified Data Observability and Agentic Data Management platform. Global enterprises use Acceldata to ensure data quality, optimize performance and costs, and build confidence in the data that powers analytics, AI, and business decisions.

To learn more, follow Acceldata on LinkedIn or on Twitter

acceldata

www.acceldata.io