Data Observability for the Modern Data Stack

A conversation with

Rohit Choudhary
Founder & CEO
Acceldata.io

Sanjeev Mohan
Principal, SanjMo & Former Gartner
Research VP, Big Data and Advanced Analytics
Moderated by **Mike Meriton**
Co-Founder & COO, EDM Council

- Joined EDM Council full-time 2015 to lead Industry Engagement
- EDM Council Co-Founder & First Chairman (2005-2007)
- Former CEO GoldenSource (2002-2015)
- Former Executive for D&B Software and Oracle
- FinTech Innovation Lab – Executive Mentor (2011 – Present)
Everyone has a data problem.

MORE DATA. MORE USE CASES. MORE COMPLEXITY

POOR PERFORMANCE
Downtime, latency, data quality and drift impact critical business operations

SLOW TO ADAPT
Data scientists, analysts, and developers struggle to identify best tech, data, and configurations for new solutions

TOO EXPENSIVE
Overprovisioning, unused data, and cost/benefit mismatches create waste
Most enterprises can’t keep up with exponentially increasing complexity of the data stack...
...Operating mission-critical data systems at scale is painful.

Adoption
- Exploding use cases
- Fast changing technology
- Limited talent pool

Operations
- Overwhelmed with scale
- Unmet SLO’s, SLA
- Low data engineering productivity

Impact

“Data lakes create sizable assets but performance is not living up to expectations”

“Data science/analytics” is the second most difficult skill set to find after “cybersecurity”

Total control of data is impossible due to the extreme distribution, complexity, and pace of change

Most analytics and AI projects fail because operationalization is an afterthought

Source: IDC, Gartner, Customer interviews, news reports
Traditional monitoring tools weren’t designed for modern data and analytics systems.

**BEFORE**

- **Apps**
- **Data Warehouse**
- **Analytics**
- **Managers**

**Purpose:** Support periodic management decisions (e.g., monthly, quarterly)

**TODAY**

- **Apps**
- **Analytics**

**Purpose:** Support operational decisions for employees, customers, suppliers, and partners in real-time
Organizations are looking for ways to simplify today’s data complexity.

**Operational Pain:**
Data operations shouldn’t feel like solving murder mysteries

**Innovation Pain:**
Skill Set Shortage

**Budget Pain:**
Fundamental tension: Reduce Risk & Reduce Cost

**Complexity:** Silo’d technologies don’t understand adjacent systems

**Support:** Coverage gaps, coordination issues

**Expectations:** Slow is the new down; quickly solving problems isn’t good enough

**Distraction:** Valuable engineering resources devoted to daily firefighting

**Frustration:** Debugging, tuning and scaling for production are tedious and time consuming

**Change:** Hard to adopt new technologies

**Provisioning:** Balancing cost and requirements fraught with risk and errors.

**Configuration:** Difficult to tune settings in constantly changing environments

**Modernization:** Migration, implementation, and management costs and issues can blow budgets
Multidimensional data observability supports the full modern data stack.
CASE STUDY: TRUE DIGITAL

True Digital scales open-source platform and saves $3m+/year.

Problem
Pervasive data system performance and scalability issues left 50% of ingested data unprocessed.

Solution
Acceldata tools isolated bottlenecks, automated performance improvements, and distinguished between mandatory and unnecessary data to ensure 8PB+ data lake could reliably support all critical enterprise analytics requirements.

Results
- Reduced annual software licensing costs by $2M+.
- Improved existing system capacity and saved additional $1M+ in projected capex.
- Eliminated unplanned outages and Sev 1 issues five consecutive months and running
- Optimized HDFS storage cost by ~2PB or 25%.

Acceldata’s tools fixed our analytics pipeline issues, improved visibility into our data systems, and recommended ways to scale and optimize our systems to meet future requirements. They helped True Digital transition to open-source technologies, allowing us to reduce licensing costs, while delivering mission-critical analytics across the enterprise.”

Wanlapa Linlawan
HEAD OF ANALYTICS PLATFORM

One of Thailand’s largest communications companies
- 100+ nodes, 8+ PB data lake cluster
- Hadoop, Hive, Spark, Ranger, Kafka open-source HDP
- ~69K Streaming messages/second
- 500M+ user impressions/month
CASE STUDY: PHONEPE (WALMART)

PhonePe scales open-source data platform by 13x, saves $5m+/year

Problem
Scaling and performance issues on open-source OLTP and OLAP.

Solution
Acceldata Pulse helped PhonePe monitor HBase, Spark, and Kafka to distinguish between infrastructure issues and seasonal and campaign-based anomalies.

Results
• Scaled data infrastructure by 13x, from 70 to 900+ nodes.
• Delivered 99.98% availability across Hadoop data lakes.
• Saved $5m+/year in software licensing costs.
• Reduced enterprise data warehouse processing time from 11.5 hours to 6.5 hours, eliminated ingestion issues, and automated 99+% reports.
• Eliminated unplanned outages and Sev 1 issues for six consecutive months and running.

Large HBase OLTP Cluster Hadoop, HBase, Hive, Spark, Ranger, Kafka
900+ Nodes, 28+ HDP Clusters, 20+ PBs
400M cash Transactions/month
1.1K Cash-transactions Per-Second

“Acceldata supports our hyper-growth and helps us manage one of the world’s largest instant payment systems. PhonePe’s biggest-ever data infrastructure initiative would never have been possible without Acceldata.”

Burzin Engineer, FOUNDER CHIEF RELIABILITY OFFICER

Acceldata supports our hyper-growth and helps us manage one of the world’s largest instant payment systems. PhonePe’s biggest-ever data infrastructure initiative would never have been possible without Acceldata.”

Burzin Engineer, FOUNDER CHIEF RELIABILITY OFFICER

Acceldata supports our hyper-growth and helps us manage one of the world’s largest instant payment systems. PhonePe’s biggest-ever data infrastructure initiative would never have been possible without Acceldata.”

Burzin Engineer, FOUNDER CHIEF RELIABILITY OFFICER
GE centralizes monitoring and boosts data system performance.

**Problem**
Unstable data infrastructure and poor performance; limited visibility into data pipelines and operations.

**Solution**
Acceldata tools provided GE’s Finance Data Lake team visibility into analytical and machine learning workloads which helped resolve data performance, user, and infrastructure issues across our systems.

**Results**
- Centralized monitoring on Acceldata, replacing multiple tools, including Grafana, Ambari, New Relic, MemSQL Ops, and Logstash.
- Supported 20+ analytics systems accessing data reports without outages or performance issues.
- Reduced Hive LLAP peak query runtime and resource costs by ~40%
- Increased engineering team productivity; enabled re-platforming consumption tier from Hadoop to SingleStore (MemSQL), while eliminating daily firefighting on outages and performance issues.

"Acceldata unifies monitoring, teams, and tools in one location so we can focus on core data-engineering. It’s a game-changer for GE. We used Acceldata to seamlessly modernize our data platforms to support data-driven decisions across our businesses."

Diwakar Goel
GLOBAL DATA OFFICER

---

10,000+ Cores
AWS Deployment

Hadoop, SingleStore (MemSQL), Spark, Greenplum
Data lake connected to 140+ ERP Systems

Consumption, ETL, Ingestion, and CDC

Acceldata — © 2021 — Confidential & Proprietary.
Questions?
Unified Data Observability Platform for the Enterprise

**FOUNDED 2018**
Delaware C-Corp
Palo Alto (HQ), Bangalore, Singapore
80+ Employees

**$45MM RAISED**
Insight, Lightspeed, Sorenson and Emergent Ventures. Some select angels.

**3X CUSTOMER GROWTH**
Observing 100 PB data
200% Net Retention

**KEY CUSTOMERS**

**ONLY COMPLETE STACK FOR DATA OBSERVABILITY**
Acceldata Observability Platform simplifies and scales data operations across all levels of compute, pipelines, and data.

**LEADERSHIP TEAM**
ACCELDATA PLATFORM

OPERATE TROUBLE-FREE
Troubleshoot, Predict and Prevent Incidents

INNOVATE FASTER
Simplify tuning for deployment at scale

OPTIMIZE RESOURCES
Right-size and configure for performance and cost

Acceldata — © 2021 — Confidential & Proprietary.
Operational intelligence to simplify complex data operations

**Multidimensional Data Observability: Advanced Analytics to Operate, Innovate & Optimize**

<table>
<thead>
<tr>
<th>Action</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predict</td>
<td>Trend Analysis</td>
</tr>
<tr>
<td>Prescribe</td>
<td>Recommendations</td>
</tr>
<tr>
<td>Prevent</td>
<td>Auto-remediation</td>
</tr>
<tr>
<td>Troubleshoot</td>
<td>Event Correlation</td>
</tr>
<tr>
<td>Optimize</td>
<td>Simulation</td>
</tr>
<tr>
<td>Contextualize</td>
<td>Data Classification, Clustering &amp; Association</td>
</tr>
<tr>
<td>Cost</td>
<td>Financial Forecasting</td>
</tr>
</tbody>
</table>
FOR MORE INFORMATION:

Loretta Jones
VP Growth
Acceldata.io
loretta@acceldata.io