High Volume Ingestion & Real Time Analytics @ Enterprise Scale

ChistaDATA

Traditional Data Platforms left Enterprises with constrains and challenges

- **Not ready for real time curation** – designed for structured Input-Output (I/O) and write intensive operations - highly tuned for batch processing
- **Limited capabilities** across critical aspects like:
  - Versatile data ingestion including real time streaming, data replication and batch processing
  - Real time analytics
  - Web scale scalability & real time data processing
- **Forced to choose Cloud or On-Prem** - lacks required capabilities for large scale Multi-Cloud and / or on prem environments, and modern API
- **Huge costs of licensing and resources** – Commercial model based on I/O and consumption. Legacy lock-in and integration issues.

How does ChistaDATA solve this problem

- ✓ **Open-Source platform** – Committed to be perennially open source driven and GPL
- ✓ **Built for Real-time analytics**
  - o Column store Database
  - o Extreme read performance
  - o Data Ingestion abilities for significant volumes
  - o Horizontal scaling
- ✓ **Extreme Automation** – For everything from deployment, troubleshooting and performance tuning
- ✓ **Proven solutions** – Already being used by some of the largest Data owners in the world

>50%
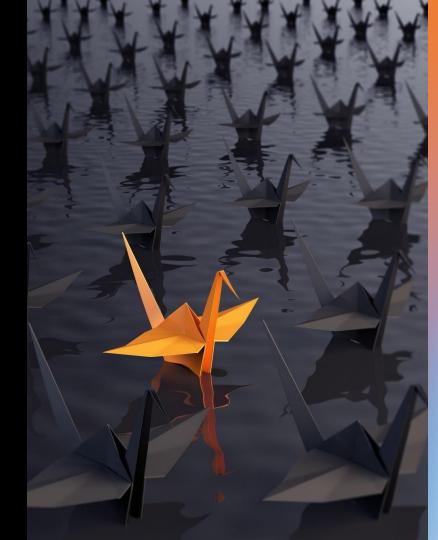Savings in Big Data / Analytics Platform TCO

50 – 200X
Faster, data processing and real time analytics
(benchmarked)

No Lock-in
Ability to be in multi-Cloud, Hybrid and On-Prem

About ChistaDATA

- **What we do**: Enabling Real-Time Analytics through extreme automation

- **How we do**: Through ClickHouse, built for Enterprise grade. On-prem and Serverless

- **What we deliver to clients:** High performance, easy migration, significant savings, On-Cloud and On-prem

- **Trusted by Global organizations** incl Paypal, Amex Travel, Garmin, Quad9, Orange Telecom and many more.

- **Strong Data lineage**

ChistaDATA – Helping clients build real-time analytics with extreme performance and significant savings

Adopted by leaders

Entrusted and adopted by market leaders like PayPal, Amex Travel Cards, Netflix, Orange Telecom, Quad9 and many others.

- Advanced and fully autonomous top class Database Infrastructure (Cloud / On-prem) for versatile high-volume ingestion, real-time Analytics and OLAP.

ClickHouse - Powered by Critical Capabilities for Digital Enterprise

01 Massive Parallel Processing
Large / complex queries can be run in parallel with minimal or no effort

02 Sophisticated Data Compression
Enabled / powered by ClickHouse for improved query performance

03 High Scalability & Availability
As a distributed system, it can seamlessly scale horizontally, also supports high availability through replication

04 Designed for Web Scale Analytics
Enriched with sharding and distributed processing. Each shard in ClickHouse can be a group of replicas addressing maximum reliability and fault tolerance.

05 Advanced Analytics
Features advanced analytical functions like window functions, aggregate functions, and SQL-based data filtering, for time-series data analysis.

06 Efficient Processing
Comes with "vectorized query execution" wherein operations are dispatched on arrays instead of individual values. Helps lowering the data processing cost.

07 Column Oriented Storage Model
Designed with Column-oriented storage model, storing data by columns allowing for efficient compression and faster data retrieval, especially for time-series data

08 Primary Key Support
Permits real-time data updates with a primary key. Data is sorted incrementally using the merge tree to perform queries on the range of primary key values.

Credentials & Capabilities that matters for your success

Empowering a Payment leader with High-Velocity, High-Volume Data Ingestion using Apache Kafka and ClickHouse

Challenges

- Lack of rapid insights for decision making
- High Complexity - Diverse data sources, including online transactions, user interactions, and operational logs
- Struggling Data Infra to cope with performance needs and high costs – lack of scalability

Solution

- **Apache Kafka Integration**: ChistaDATA implemented a robust data streaming pipeline using Apache Kafka.
- **ClickHouse Deployment**: ChistaDATA helped set up a ClickHouse cluster optimized for real-time analytics.
- **Data Transformation**: Implemented data transformation and enrichment processes within the Kafka pipeline, enabling preprocessing and enrich incoming data before storage.
- **Real-Time Dashboards**: Using ClickHouse's integration with visualization tools. These dashboards provided instant insights into payment trends, user behaviors, and operational performance.

Benefits

- Real-Time Insights
- Scalability without compromised performance
- Enhanced accuracy and reduced data inconsistencies
- Operational efficiency and cost savings

Also delivering differentiation on the wider spectrum of real time data processing and advanced analytics

+12,00.50

	Challenge	Solution	Benefits
Large Telco	1. Rise in fraudulent activities 2. Analyze large volumes of CDRs in real-time. 3. Traditional DBs unable to manage	1. ClickHouse cluster to ingest and store CDRs in real-time 2. Execute complex queries and Query optimization 3. Interactive Dashboards and alerts	✓ Significant savings of > 50% ✓ Much higher performance – faster speeds & real-time ✓ Minimized losses to business
Global Ad-tech	1. Lack of real-time insights for timely ad campaigns 2. Price, performance & scalability constraints	1. Data ingestion and ETL pipeline for data streaming 2. Migration from Snowflake 3. Data model & schema created and deployed 4. Query optimization	✓ Faster performance ✓ Extreme cost savings ✓ Improved insights, enabling real-time targeted ad campaigns

We offer a long term solution which is future proof and easy to adopt

Forever open-source and GPL

Built for significant volume and speed

Proven – Adopted by Data intensive leaders

Horizontal Scaling and Performance assurance

Capabilities matching your needs

Key Platform Features	Data Ingestion Variants	Data Security Requirements	Operational Analytics Dashboard
☑ Self-service platform ☑ Scale-out architecture ☑ Elastic Extraction, processing, and storage ☑ Built on native PaaS ☑ Feed logging / monitor / audit ⇔ Dashboard capability with multiple views ⇔ Job scheduling and orchestration ⇔ Operational metrics	☑ Batch ☑ API and micro-batch ☑ Real-time streaming ⇔ Extensible out of box	☑ Data Masking ☑ Create / Update / Execute / View Data Encryption policy ☑ Add / Update / Delete ☑ Advanced multi-level data encryption for data at rest as well as in transit ☑ Scheduled and custom backups capability available	☑ Minimum / Maximum / Average Data Record Count ☑ No. of data items loaded / success / failures ⇔ Duration of Feeds

Data Management	Data Entitlements	Data Controls	Processing Controls
☑ Data Feed Repository ☑ Metadata extraction and management ☑ Add / View / Extract metadata ⇔ Metadata generation ⇔ Data lineage ⇔ End-end traceability view of data starting from the origin ⇔ Feed metadata registration interface from the dashboard	☑ Access controlled usage of data sets ☑ Allow privileged access / prohibits data access for roles / individuals ☑ Data transformation and enrichment capability ☑ User and access management ⇔ Business Glossary ⇔ Create / view / tag Business catalogue ⇔ Data Profiling ⇔ Data Scanning ⇔ PII and sensitive data detection capability	☑ Processing Boundary ☑ End-end audit trails ☑ All objects analytics ☑ Data integrity measurement / validation, including audit controls ⇔ Automated capability for incomplete / corrupt / missing data identification ⇔ Alert & notification triggering capability ⇔ Metadata altering	☑ Capacity Management ☑ Reactive and proactive monitoring / expansion ⇔ Definition and management of KPIs and metrics ⇔ Automated dashboard for data and processing controls management ⇔ Data Quality Rules

THANK YOU

ChistaDATA