WANdisco LiveAnalytics and Managed Delta Lake on Databricks

Businesses are under pressure to decrease their mean time to insight in order to accelerate growth. They are driving to better understand markets, refine operations and create new business models.

The move to cloud analytics is making companies more competitive, lean, and nimble. Businesses are taking advantage of cheap, scalable storage and the flexibility that the cloud offers for modern analytics.

This is putting the Hadoop market under pressure. Legacy Hadoop customers are moving workloads into public clouds. Since the merger of the two largest Hadoop vendors there has been increased demand for cloud analytic solutions as customers pause expansion of Hadoop usage and consider their cloud options.

At the same time, hardware maturity cycles are prompting customers to avoid an expensive refresh of capital investments and instead opting for operational expenditure in the cloud. Analytical capabilities in the cloud have surpassed those of on-premises Hadoop installations with the broad availability of technologies like Databricks.

How to minimize risk to take advantage of the significant economic and operational benefits of cloud analytics?

Business success is dependent on analyzing petabytes of data to accelerate outcomes and create competitive advantage. While the move to cloud analytics is making companies more competitive, lean, and nimble, the transformation is fraught with risk of disrupting the existing business. The WANdisco's LiveAnalytics and Databricks joint solution ensures enterprise data in the cloud is always available, accurate and protected. LiveAnalytics helps you:

1. **Automate data migration**
   with Zero Downtime, Zero Data Loss
2. **Ensure the reliability and consistency of data**
   with Databricks Unified Analytics Platform
3. **Replicate changes immediately**
   from Hive content on premises to equivalent changes in the cloud environment

Challenges moving analytics to the cloud

With such compelling advantages, timing and market forces for big data in the cloud, what prevents organizations from adopting it today? We see common impediments to cloud adoption for big data:

Sunk cost: Enterprises have heavy investment in existing skills, infrastructure and systems built using on-premises Hadoop Infrastructure

Data gravity: In order to use all enterprise data with reasonable efficiency and speed, applications have historically needed to be close to the source for higher throughput and lower latency

Strict SLAs: Enterprises who have big data environments must move to the cloud without business disruption

The WANdisco and Databricks joint solution helps businesses overcome these challenges, cut operational cost, meet SLAs, improve data reliability and increase analytics capabilities by automating the movement of large scale on-premises data to cloud

About WANdisco

With zero downtime and zero data loss, WANdisco LiveMigrator solves the problem of moving petabyte scale data to the cloud and LiveAnalytics Solution provides immediate analytic data access through continuous automated replication from on-premises Hadoop analytics to Spark based cloud analytics. WANdisco Fusion keeps geographically dispersed data at any scale consistent between on-premises and cloud environments allowing businesses to operate seamlessly in a hybrid or multi-cloud environment.

WANdisco has over a hundred customers and significant go-to-market partnerships with Microsoft Azure, Amazon Web Services, Google Cloud, Oracle, and others as well as OEM relationships with IBM and Alibaba.

1. AUTOMATE DATA MIGRATION

Migrate from on-premises HDFS to Managed Delta Lake on Databricks running in Azure

Migration is the first step the cloud analytics transformation journey. One common migration approach, Manual Cloud Migration, is based on custom programming to copy data to the cloud. Manual cloud migration includes incomplete strategies such as change data capture, which has no visibility of existing data, only of changes, tools to copy static data such as DistCp, which do not account for changes being made to the data under migration, or dual-ingest architectures which require that your applications are modified significantly.

Large data sets take time to bring to the cloud. While making data available in the cloud, change and ingest is still needed. This time lag makes it challenging to accurately bring continuously changing large scale data sets to the cloud. Manual migrations create unnecessary business risk because they can disrupt the operation of on-premises applications, deliver inconsistent data, risk data loss, and lack accuracy validation. In addition, they cause costly overhead with time overruns, manual checking, repeated scans, and significant resources for custom code maintenance and management.

LiveAnalytics Automated Migration

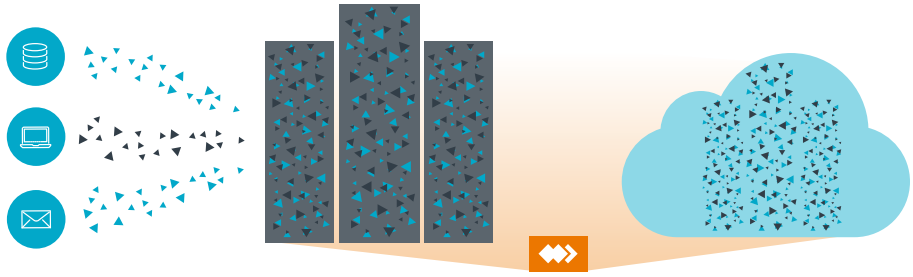

Figure 1: WANdisco LiveAnalytics Solution Automates Cloud Migration

Does your need for business continuity prevent you moving to the cloud?

If you have business-critical systems running on your big data environments, moving them away from cumbersome and costly fixed infrastructure can be challenging.

You want to take advantage of the cloud, where access to modern data and processing capabilities is more flexible, but getting there may mean stopping your systems while you migrate.

You need a solution that can let you maintain business operations while you migrate massive volumes of data efficiently.

WANdisco LiveAnalytics Solution automates your Hadoop data to Delta Lake migration at scale with no application downtime and no risk of data loss, even when your data sets are under active change.

Migrate at scale from continuously operating on-premises systems to cloud storage and multiple cloud regions.

As changes occur anywhere in the donor system, LiveMigrator creates and validates beneficiary data consistency.

LiveAnalytics Solution minimizes IT resources with One Click replication across all major commercial Hadoop distributions, cloud storage and analytic services, while requiring just One Pass of the source storage.

- Migrates petabyte-scale big data sets to cloud storage without needing to halt changes made to the data sets during migration.
- Uses an efficient, One Pass approach to scan data.
- Applies WANdisco's patented Fusion technology to ensure that ongoing changes are applied to the target migration.
- Transitions seamlessly to a hybrid architecture to support extended periods of validation for applications and consumers of the cloud storage.
- Eliminates points of failure by recovering from network outages and disruptions automatically and without user intervention.
- Minimizes the time required to bring workloads to the cloud by making each data location available for use as soon as bandwidth allows.
- Works across a variety of big data source and target environments, including all major Hadoop and object storage technologies.

2. ENSURE THE RELIABILITY AND CONSISTENCY OF DATA

With unified analytics on one platform

While the cloud brings efficiencies for data lakes there remains concerns about the reliability and the consistency of the data. Data Lakes typically have multiple data pipelines reading and writing data concurrently, and data engineers have to go through a tedious process to ensure data integrity, due to the lack of transactions.

Databricks handles with ease all of the analytic processes that previously suffered under inflexible and cumbersome Hadoop deployments on-premises.

Its success is prompting enterprises like yourselves to consider migrating on-premises Hadoop workloads to the cloud and in particular to Databricks. Databricks provides an elegant answer to the concern around the sunk costs of skills, infrastructure and systems built on-premises in Hadoop by allowing the same technologies, applications and systems to operate without change either on-premises or in the cloud. Delta Lake provides the storage layer on top of your existing storage to support enterprise workloads across streaming and batch requirements to better manage data lakes at scale. Delta Lake supports multiple simultaneous readers and writers for mixed batch and streaming data making it easy for data teams to run interactive queries and batch historic backfill out of the box.

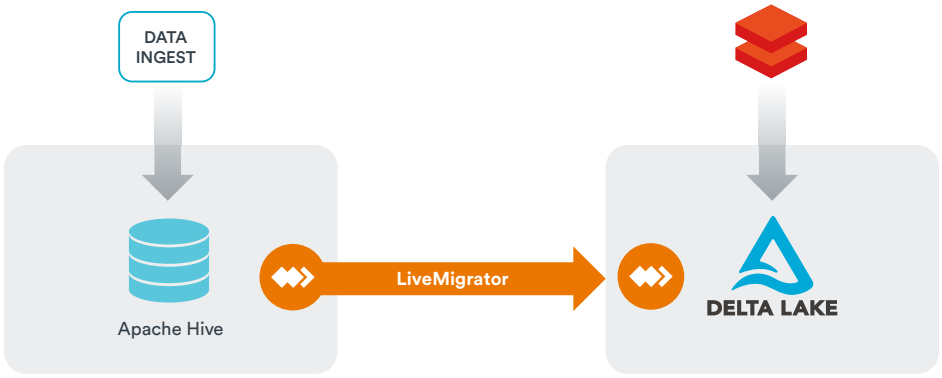

Figure 2: LiveMigrator and Delta Lake

Databricks provides a Unified Analytics Platform powered by Apache Spark for data science teams to collaborate with data engineering and lines of business to build data products. You can achieve faster time-to-value with Databricks by creating analytic workflows that go from ETL and interactive exploration to production. Databricks also makes it easier for you to focus on your data rather than hardware by providing a fully managed, scalable, and secure cloud infrastructure that reduces operational complexity and total cost of ownership.

Delta Lake brings key features to cloud storage that have challenged adopters of the cloud:

ACID Transactions
Delta Lake brings ACID transactions to your data lakes. It provides serializability, the strongest level of isolation level, allowing you to build reliability into your data processing and analytics effortlessly.

Scalable Metadata
Delta Lake treats metadata just like data, leveraging Spark's distributed processing power to handle all its metadata. As a result, Delta Lake can handle petabyte-scale tables with billions of partitions and files at ease.

Data Versioning
Delta Lake provides snapshots of data enabling developers to access and revert to earlier versions of data for audits, rollbacks or to reproduce experiments.

An Open Data Format
All data in Delta Lake is stored in Apache Parquet format enabling Delta Lake to leverage the efficient compression and encoding schemes that are native to Parquet.

Unified Batch and Streaming
Tables in Delta Lake support batch and streaming interactions. Streaming data ingest, batch historic backfill, and interactive queries work directly.

Schema Management
Delta Lake can enforce defined schemas to ensure that data types are correct and required columns are present, preventing bad data from causing data corruption.

Schema Evolution
Big data is continuously changing. Delta Lake applies changes to table schema automatically, without the need for cumbersome DDL.

Apache Spark Compatibility
Developers can use Delta Lake with their existing data pipelines with minimal change as it is fully compatible with Spark.

3. REPLICATE CHANGES IMMEDIATELY

From Hive content on premises to equivalent changes in the cloud environment

With WANdisco LiveMigrator we solved the problem of moving petabyte scale data to the cloud without stopping your business. With LiveAnalytics, WANdisco provides a continuous replication solution from on-premises Hadoop analytics to Spark based cloud analytics with zero downtime and zero data loss.

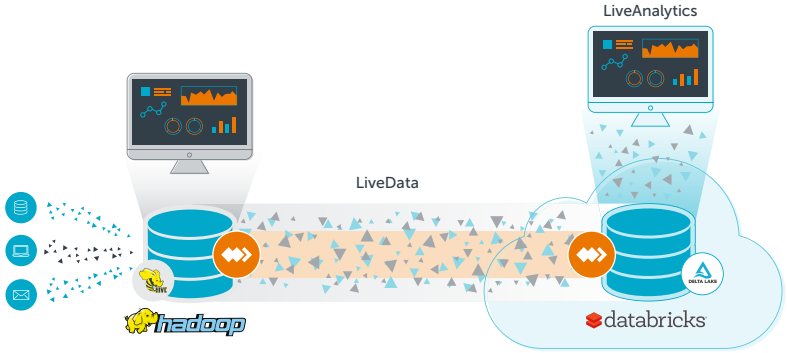

Figure 3: LiveAnalytics provides continuous replication from on-premises Hadoop analytics to Spark-based cloud analytics

The LiveAnalytics Solution

The combination of WANdisco LiveAnalytics and managed Delta Lake on Databricks provides an automated risk free answer to the significant challenges of migrating enterprise big data systems to what is clearly a better place: the cloud. More specifically, they provide a solution for organizations to take advantage of Databricks Unified Analytics Platform in the cloud without disrupting their business operations just to do so.

WANdisco uses a unique distributed coordination engine to guarantee that changes made to replicas of data are available among every instance, meaning that every user and application has read and write access to the data they need, regardless of geographic location, underlying storage platform or cloud storage provider. Any changes made to Hive content on premises, whether that is ingesting data, deleting data, modifying schema information, is immediately reflected with equivalent changes in the cloud environment.

LiveAnalytics' continuous, consistent, automated data replication ensures migrated data is immediately available for analytical processing in managed Delta Lake on Databricks. With minimal disruption when migrating between Hadoop and non Hadoop environments, LiveAnalytics and Databricks provide faster adoption of ML and AI for enterprises. LiveAnalytics keeps your data accurate and consistent across all your business application environments, regardless of geographic location, data platform architecture, or cloud storage provider.

What is LiveAnalytics Solution?

One solution to ensure data migrates and stays accurate and consistent across business application environments, regardless of geographic location, data platform architecture, or cloud storage provider.

LiveMigrator

- Provides **One Click** automated replication across all major commercial Hadoop distributions, cloud storage and analytic services
- Requires just **One Pass** of the source storage
- Delivers **Validated Data Consistency**

Fusion

- Enterprise-class software platform
- Ensures Data Consistency
- Keeps data consistent in a distributed environment – on-premises, hybrid-cloud, multi-region cloud, and multi-cloud

Plugin for Delta Lake

- Replicates Hive content to cloud storage for Databricks
- Makes Hive data and metadata available as Delta Lake tables

5000 Executive Parkway, Suite 270
San Ramon, CA 94583

www.wandisco.com

Talk to one of our specialists today

US	+1 877 WANDISCO (926-3472)
EMEA	+44 (0) 114 3039985
APAC	+61 2 8211 0620
All other	+1 925 380 1728

Join us online to access our extensive resource library and view our webinars.

Follow us to stay in touch