Modern Data Architecture with Apache™ Hadoop®
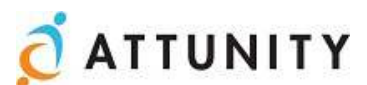
Automating Data Transfer with Attunity Replicate

Presented by Hortonworks and Attunity

Hortonworks

ATTUNITY

Executive Summary

Apache Hadoop didn't disrupt the datacenter, the data did.

Shortly after corporate IT functions within enterprises adopted large-scale systems to manage data, the Enterprise Data Warehouse (EDW) emerged as the logical home of all enterprise data. Today, every enterprise has a data warehouse that serves to model and capture the essence of the business from their enterprise systems.

The explosion of new types of data in recent years – from inputs such as the web and connected devices, or just sheer volumes of records – has put tremendous pressure on the EDW.

In response to this disruption, an increasing number of organizations have turned to Apache Hadoop to help manage the enormous increase in data while maintaining coherence of the data warehouse.

This paper discusses Apache Hadoop, its capabilities as a data platform and how one can take advantage of Attunity Replicate, a high-speed data loading and change data capture (CDC) solution, to accelerate the adoption of Hadoop by reducing the complexity of moving Big Data to and from the technology platform.

Apache Hadoop provides these benefits through a technology core comprised of:

Hadoop Distributed Filesystem (HDFS) - a Java-based file system that provides scalable and reliable data storage designed to span large clusters of commodity servers.

Apache Hadoop YARN - provides a pluggable architecture and resource management for data processing engines to interact with data stored in HDFS.

Attunity Replicate automates data transfer into and out of Hadoop and the enterprise data lake from many heterogeneous data sources. Using Attunity Replicate, organizations can achieve faster time-to-value for Big Data projects and deliver a more complete and trusted view of the business. As a result, companies can harness the power of Big Data to drive new insights and deliver competitive advantage.

The Attunity solution for Hadoop consists of:

Attunity Replicate - an optimized and automated platform for moving structured data into and out of Hadoop including data from all major databases, data warehouses, and structured files

Attunity Maestro - a highly scalable workflow management platform that orchestrates and automates data transmission and deployment processes of Big Data, applications, and large file assets.

For an independent analysis of Hortonworks Data Platform, download [Forrester Wave™: Big Data Hadoop Solutions, Q1 2014](#) from Forrester Research.

To learn more about moving all data types in and out of Hadoop, download the [Hadoop and the Modern Data Supply Chain](#) whitepaper from CITO Research.

Disruption in the Data

Corporate IT functions within enterprises have been tackling data challenges at scale for many years now. The vast majority of data produced within the enterprise stems from large-scale Enterprise Resource Planning (ERP) systems, Customer Relationship Management (CRM) systems, and other systems supporting a given enterprise function. Shortly after these "systems of record" became the way to do business, the EDW emerged as the logical home of data extracted from these systems to unlock "business intelligence" applications, and an industry was born. Today, every organization has data warehouses that serve to model and capture the essence of the business from their enterprise systems.

The Challenge of New Types of Data

The emergence and explosion of new types of data in recent years has put tremendous pressure on all of the data systems within the enterprise. These new types of data stem from "systems of engagement" such as websites, or from the growth in connected devices.

The data from these sources has a number of attributes that make it a challenge for a data warehouse:

Exponential Growth. An estimated 2.8ZB of data in 2012 is expected to grow to 40ZB by 2020. Eighty-five percent of this data growth is expected to come from new types, with machine-generated data being projected to increase 15x by 2020. (Source: IDC)

Varied Nature. The incoming data can have little or no structure, or structure that changes too frequently for reliable schema creation at time of ingest.

Value at High Volumes. The incoming data can have little or no value as individual or small groups of records. But at high volumes or with a longer historical perspective, data can be inspected for patterns and used for advanced analytic applications.

The larger scale in data volumes and number of sources represents a challenge in cost, time and efficiency. This raises the need for technologies that can accelerate the process of moving data, and enable efficiencies in implementing new data feeds as well as optimizing high performance data transfer.

The Growth of Apache Hadoop

Challenges of capture and storage aside, the blending of existing enterprise data with new types of data is being proven by many enterprises across virtually all industries from retail, financial services, and healthcare to advertising, manufacturing and energy.

The technology that has emerged as the way to tackle the challenge and realize the value in Big Data is Apache Hadoop, whose momentum was described as "unstoppable" by Forrester Research in the *Forrester Wave™: Big Data Hadoop Solutions, Q1 2014*.

The maturation of Apache Hadoop in recent years has broadened its capabilities from simple data processing of large data sets to a full-fledged data platform with necessary services for the enterprise -- from security to operations management and more.

Find out more about these new types of data at
Hortonworks.com
• Clickstream
• Social Media
• Server Logs
• Geolocation
• Machine and Sensor

What is Hadoop?
Apache Hadoop is an open-source technology born out of the experience of web scale consumer companies such as Yahoo, Facebook and others, who were among the first to confront the need to store and process massive quantities of digital data.

Moving Data Into and Out of Hadoop

The New York Times reported that data scientists may spend as much as 80 percent of their time on "data wrangling" collecting and moving large volumes of data before it can be explored for useful nuggets.[1] This is because Big Data is heterogeneous by nature, so data analysts need to find fast and cost-effective ways to move information from many different sources into the Hadoop system and to analyze it before it becomes irrelevant.

CITO Research writes that enterprises should implement solutions that are specifically designed to ease and accelerate the process of data movement across a broad number of platforms.[2] Their research goes on to suggest that these technologies need to empower IT organizations to easily move data from one depository to another in a highly visible manner. Effective solutions should also unify and integrate data from all platforms within an enterprise, not just Hadoop. And, they should include change data capture (CDC) technology so that after initial data replication, the changed data can be captured and applied in order to keep the target data up-to-date. With data movement software, enterprises can not only unleash the full power of Hadoop, but also unleash the full power of their other technologies.

When Moving Data, Bottlenecks Can Result in:

- *Failed execution of business critical Big Data projects*
- *Inability to create data lakes to support high-level analytics*
- *Limited view of the business*
- *Stale data that has lost relevance and value*

---

[1] "*For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights*," by Steve Lohr, New York Times, August 17, 2014.
[2] "*Hadoop and the Modern Supply Chain*," CITO Research, 2014.

Hadoop and Your Existing Data Systems:
A Modern Data Architecture

From an architectural perspective, the use of Hadoop as a complement to existing data systems is extremely compelling: an open source technology designed to run on large numbers of commodity servers. Hadoop provides a low-cost, scale-out approach to data storage and processing and is proven to scale to the needs of the very largest web properties in the world.
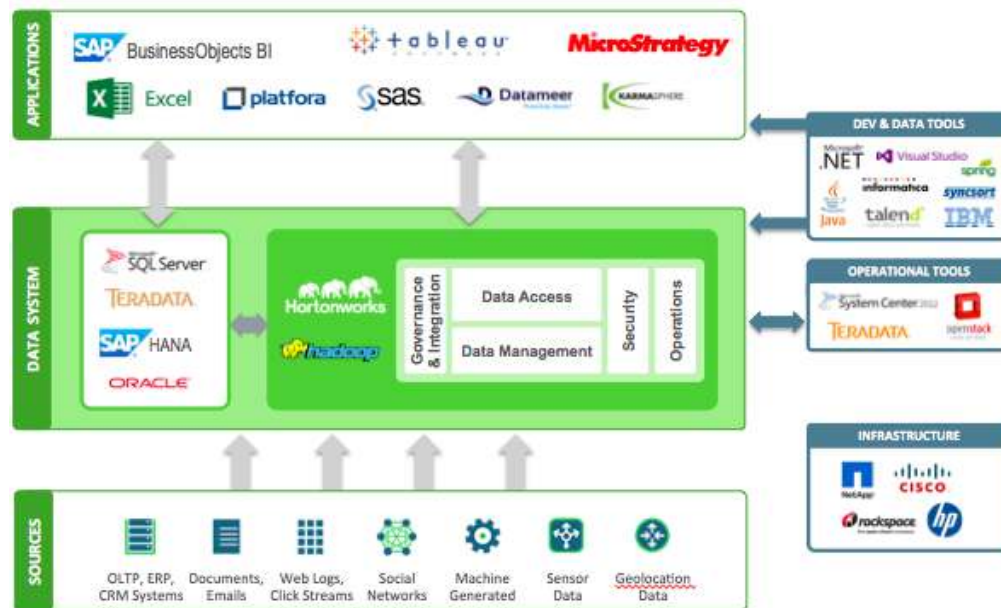

Fig. 1
A Modern Data Architecture with Apache Hadoop integrated into existing data systems

Hortonworks is dedicated to enabling Hadoop as a key component of the data center, and having partnered closely with some of the largest data warehouse vendors, it has observed several key opportunities and efficiencies that Hadoop brings to the enterprise. By combining Hortonworks' Apache Hadoop expertise with Attunity's data integration technology, enterprises can now move massive amounts of data to and from Hadoop.

New Opportunities for Analytics

The architecture of Hadoop offers new opportunities for data analytics:

Schema On Read. Unlike an EDW, in which data is transformed into a specified schema when it is loaded into the data warehouse – requiring "Schema On Write" – Hadoop empowers users to store data in its raw form and then analysts can create the schema to suit the needs of their application at the time they choose to analyze the data, empowering "Schema On Read." This overcomes issues around the lack of structure and investing in data processing when there is unknown initial value of incoming data.

> Attunity offers a 'Click-2-Replicate' solution that delivers 'raw' data into Hadoop without any development or scripting efforts.

Multi-use, Multi-workload Data Processing. By supporting multiple access methods (batch, real-time, streaming, in-memory, etc.) to a common data set, Hadoop enables analysts to transform and view data in multiple ways (across various schemas) to obtain closed-loop analytics by bringing time-to-insight closer to real time than ever before.

> Attunity provides many data loading and synchronization options that enable various workloads in Hadoop.

New Efficiencies for Data Architecture

In addition to the opportunities for Big Data analytics, Hadoop offers efficiencies in a data architecture:

Lower Cost of Storage. By design, Hadoop runs on low-cost commodity servers and direct-attached storage that allows for a dramatically lower overall cost of storage. In particular, when compared to high-end Storage Area Networks (SAN) from vendors such as EMC, the option of scale-out commodity compute and storage using Hadoop provides a compelling alternative–and one that allows the user to scale-out their hardware only as their data needs grow. This cost dynamic makes it possible to store, process, analyze, and access more data than ever before.

> Attunity enables companies to migrate and move data from high cost data management and storage systems Into Hadoop.

Data Warehouse Workload Optimization. The scope of tasks being executed by the EDW has grown considerably across ETL, analytics and operations. However, the transformation process (the "T" in "ETL") is a relatively low-value computing workload that can be performed on in a much lower-cost manner. Many users offload this function to Hadoop, wherein data is transformed and then the results are loaded into the data warehouse.

> Attunity delivers efficient and easy-to-use "E" and "L" from the EDW into and out of Hadoop. With transformations completed in Hadoop, this reduces unnecessarily bundled ETL costs.

The result: critical CPU cycles and storage space can be freed up from the data warehouse, enabling it to perform the truly high-value functions –analytics and operations –that best leverage its advanced capabilities.

Enterprise Hadoop with Hortonworks Data Platform

To realize the value in your Big Data investment, use the blueprint for Enterprise Hadoop to integrate with your EDW and related data systems. Building a modern data architecture enables your organization to store and analyze the data most important to your business. At massive scale, enterprises can extract critical business insights from all types of data from any source, and ultimately improve your their competitive position in the market and maximize customer loyalty and revenues. Read more at *http://hortonworks.com/hdp*.

Hortonworks Data Platform is the foundation for a Modern Data Architecture

Hortonworks Data Platform (HDP) is powered 100% by Open Source Apache Hadoop. HDP provides all of the Apache Hadoop-related projects necessary to integrate Hadoop alongside an EDW as part of a Modern Data Architecture.
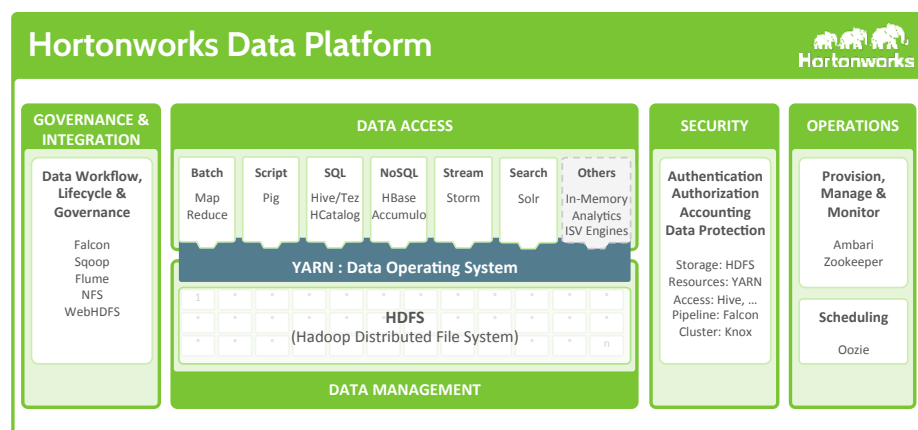

Fig. 2

Data Management: Hadoop Distributed File System (HDFS) is the core technology for the efficient scale-out storage layer, and is designed to run across low-cost commodity hardware. Apache Hadoop YARN is the prerequisite for Enterprise Hadoop as it provides the resource management and pluggable architecture for enabling a wide variety of data access methods to operate on data stored in Hadoop with predictable performance and service levels.

Data Access: Apache Hive is the most widely adopted data access technology, though there are many specialized engines. For instance, Apache Pig provides scripting capabilities, Apache Storm offers real-time processing, Apache HBase offers columnar NoSQL storage and Apache Accumulo offers cell-level access control. All of these engines can work across one set of data and resources thanks to YARN. YARN also provides flexibility for new and emerging data access methods, including search and programming frameworks such as Cascading.

Data Governance & Integration: Apache Falcon provides policy-based workflows for governance, while Apache Flume and Sqoop enable easy data ingestion, as do the NFS and WebHDFS interfaces to HDFS.

Security: Security is provided at every layer of the Hadoop stack from HDFS and YARN to Hive and the other Data Access components on up through the entire perimeter of the cluster via Apache Knox.

Operations: Apache Ambari offers the necessary interface and APIs to provision, manage and monitor Hadoop clusters and integrate with other management console software.

Deployment Options for Hadoop

HDP offers multiple deployment options:

On-premises: HDP is the only Hadoop platform that works across Linux and Windows.

Cloud: HDP can be run as part of IaaS, and also powers Rackspace's Big Data Cloud, and Microsoft's HDInsight Service, CSC and many others.

Appliance: HDP runs on commodity hardware by default, and can also be purchased as an appliance from Teradata.

The Attunity Solution for Hadoop

Attunity delivers a high-performance solution that moves Big Data into and out of Hadoop with speed and ease. Attunity Replicate helps to access and load massive amounts of data for analytics in the cloud and datacenter. And, Attunity Maestro orchestrates and automates data transmission and deployment processes of Big Data, applications, and large file assets.
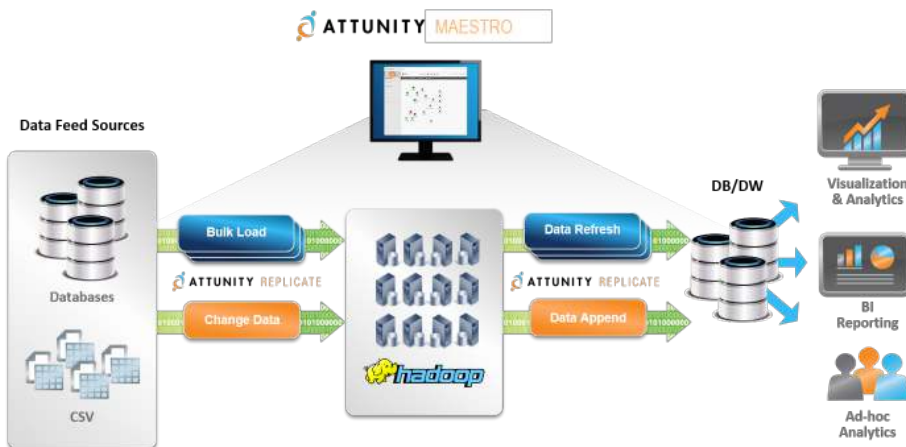

Fig. 3 Attunity Solution for Hadoop

The Attunity solution provides high-speed connectivity for collecting data out of most enterprise data sources. This data is then automatically loaded into Hadoop where it is made available for any Hadoop application to work on. Users who are tasked with bringing data into Hadoop using Attunity Replicate do not need to learn Hadoop to perform this task. This reduces the need for additional Hadoop training and personnel - maximizing its full potential by tapping into any data source and target as well as avoiding additional investments in training and hiring. The intuitive and industry-proven Attunity Replicate includes a GUI that enables users to accelerate Hadoop installations for any data delivery project. No specialized coding is required to scale performance on distributed computing platforms.

Key Features/Benefits are:

- High-performance connectivity to Hadoop though native APIs for data ingest and publication
- Automated schema generation in HCatalog
- Drag & drop configuration with "Click-2-Replicate" design
- High-speed data load options:
  - Full reload with overwrite
  - Insert only appends
  - Change Data Capture

- On-the-fly data filtering and transformation
- Compression: Gzip
- Monitoring dashboard with web-based metrics, alerts and log file management

Using Attunity's solution for Hadoop, enterprises can:

- Reduce time and resources required to move data for Hadoop
- Lower the costs associated with moving data for Hadoop
- Move data in batch as well as incrementally with low latency
- Automate movement across Hadoop and data warehouses
- Use Hadoop as both a source and a target system
- Manage the data supply chain - including data lakes - through a visual user interface

Use Case

Enterprises using Attunity Replicate to bring data into Hadoop don't need to learn Hadoop to perform this task. From managed care providers and telecom companies to major online travel reservation firms and media companies, enterprises are finding that Attunity Replicate is the easiest solution to set up and use and that it provides remarkable speed and automated data collection and loading into Apache Hadoop.

Here's how one managed care provider used Attunity Replicate in their Hadoop environment:

A provider of managed care services targeted to government-sponsored health care programs serving approximately 4 million members in North America, created a data lake leveraging Hadoop technology. The data lake serves as a repository for storing and processing data before sending some of the data to their primary analytical platform, a Pivotal data warehouse.

As an existing Attunity Replicate customer, the managed care provider understood the benefits of using Attunity Replicate to move data from Oracle and SQL Server OLTP systems into Pivotal. So, when the company decided to implement a data lake between those systems, they chose Attunity's technology to move their data in the Hortonworks Hadoop environment as well.

Other common use cases that require Hadoop data be available when, where and how it's needed are:

- ETL offload to Hadoop
- Long-term data archiving to Hadoop while online data is stored and continuously refreshed in the enterprise data warehouse
- Data delivery to a structured operational data store for business intelligence and analytics

Conclusion

Big Data has changed the way that we use and manage data. We now have more data than we've ever had before in higher velocities from more sources across the organization. Enterprises can't afford to miss business opportunities due to time spent "data wrangling" in order to mine their data for useful nuggets. Working together, Attunity and Hortonworks offer a solution to alleviate those challenges. Enterprises are leveraging the joint solution today to dramatically improve the flow and accessibility of Big Data to achieve faster time-to-value and competitive advantage.

About Attunity

Attunity is a leading provider of information availability software solutions that enable access, management, sharing and distribution of data, including Big Data, across heterogeneous enterprise platforms, organizations, and the cloud. The company's software solutions include data replication, data flow management, test data management, change data capture, data connectivity, enterprise file replication (EFR), managed-file-transfer (MFT), and cloud data delivery. Attunity has supplied innovative software solutions to its enterprise-class customers for nearly 20 years with successful deployments at thousands of organizations worldwide. For more information, visit www.attunity.com.

About Hortonworks

Hortonworks develops, distributes and supports the only 100% open source Apache Hadoop data platform. Our team comprises the largest contingent of builders and architects within the Hadoop ecosystem who represent and lead the broader enterprise requirements within these communities. The Hortonworks Data Platform provides an open platform that deeply integrates with existing IT investments and upon which enterprises can build and deploy Hadoop-based applications. Hortonworks has deep relationships with the key strategic data center partners that enable our customers to unlock the broadest opportunities from Hadoop. For more information, visit http://www.hortonworks.com.