Technology Insight Paper

Hadoop on Shared, Software-defined Storage

Using IBM Spectrum Scale with Hortonwork's Data Platform to accelerate time to information and add value to Hadoop's storage envirnment

By John Webster

June, 2017

Evaluator Group

Enabling you to make the best technology decisions

Evaluator Group

Data is the New Equity

Enterprise CEOs and other organizational leaders see new business opportunities in the rising tide of data that now surrounds them. Consumers become data sources as they employ a growing number of ways to connect with information and with others. New sources of “machine-generated” data—typically the unseen data produced by wired and wireless devices—appear every day. Leveraging the growing list of data sources is now a top IT priority. With the advance of data analytics, executives now realize that data owned by the enterprise can be monetized. Data is the new equity.

Data analysis in the enterprise has advanced from data warehouses that delivered weekly, monthly and year-end reports to real time information systems that build revenue and enhance operational efficiency and security. Big Data encompasses a set of technologies in current use for large scale data analytics with low latency and at relatively low cost. In 2017, enterprises will progress from experimentation to the wide usage of these technologies to power specific business initiatives within all major industry segments. Examples include:

Industry 4.0

Manufacturing is once again being revolutionized by a convergence of technologies and data sources. Computational power unites new Big Data analytics capabilities with wireless sensors, new human-machine interaction technologies (touch interfaces and augmented reality systems), 3D printing, and advanced robotics. Industry 4.0 initiatives yield significant advances in manufacturing efficiency, product quality, and employee safety.

Customer 360

Many CEOs now believe that they can gain or at least maintain a competitive advantage from a deep understanding of the customer. They want to capture and store as much data about their customers as possible and bring it to bear whenever the customer interacts with the business. This they believe will develop customer loyalty, boost satisfaction and increase revenue as a result.

Insurtech

The insurance industry wants to leverage Big Data analytics improve operations across a wide range of processes from sales to underwriting. Real-time/near real-time data acquisition and processing using wireless sensors, wearables, smart phones and other connected devices and wearables — will allow insurers to better manage risk, improve subscriber loyalty and optimize sales opportunities.

Proximity Marketing

Retailers have long wanted to imitate the on-line shopping experience for in-store consumers. Wireless beacons strategically situated in shopping areas are an enabling technology now being exploited by forward-leaning retailers. Customers interact with the store in real time via simple and intuitive smartphone apps that allow them to browse and buy while the retailer automates targeted commercial content delivery and collects essential data about in-store consumer behavior.

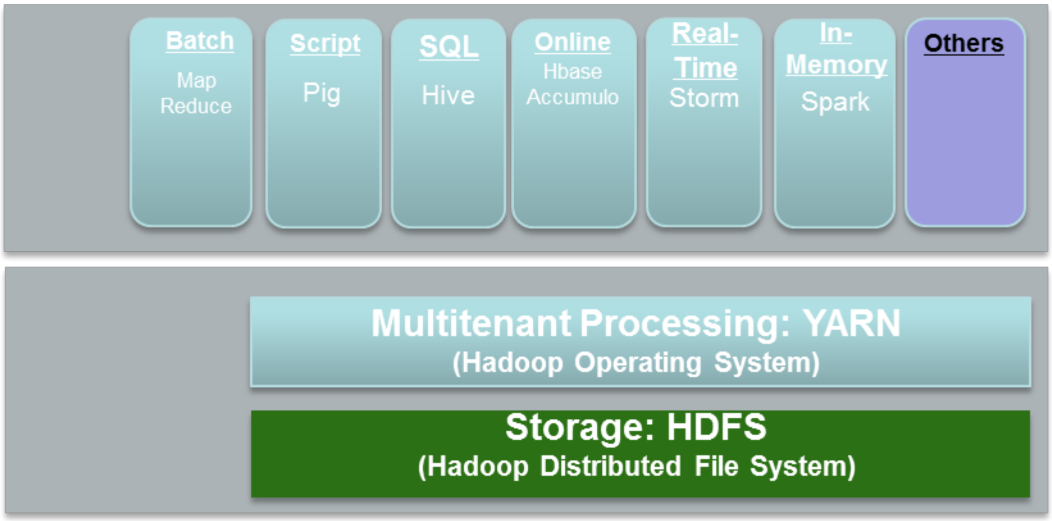
Data Monetization

CEOs are aware that their enterprise IT departments hold data that could be sold or “rented.” Through initiatives like the Internet of Things (IoT) and others like those mentioned above, they will also be gathering data that could be monetized. The classic example is that of a large automobile manufacturer gathering data via sensory devices and other data sources embedded in cars. This data could be of significant value to supply chain partners, dealers, and other related industries like tire manufacturing.

Hadoop and the Case for External Storage

Chief among today’s Big Data analytics platforms being considered for these new business initiatives is Apache Hadoop. The open source project called Apache Hadoop originated from the Internet data centers of Google and Yahoo! that are known for very large scale at low cost per unit of compute power. Hadoop offers distributed processing of large unstructured data sets. Its native storage environment—the Hadoop Distributed File System (HDFS)—is a parallelized, distributed, java-based filesystem designed for use in Hadoop clusters that currently scales to 200 PB and can support single Hadoop clusters of 4,000 nodes. HDFS supports simultaneous data access from multiple applications and users using Apache YARN (Yet Another Resource Negotiator). Hadoop is also designed to be fault tolerant meaning that it can withstand disk and node failures without interruption to cluster availability. Failed disks and cluster nodes can be replaced as needed.

Early-on, users leveraged Hadoop’s MapReduce engine to run analytics applications in batch processing mode. Now, thousands of enterprises view Big Data analytics capabilities as critical enablers of future business initiatives such as those mentioned above, and want to use Hadoop for a wide range of analytics applications. In addition to running successive MapReduce jobs within the same cluster, they also want to host multiple applications for multiple types of analytics users (see Figure 1 below). These include OLTP (Hbase) and real time analytics (Storm and Spark).



The Hadoop multi-application processing environment (Source: Hortonworks)

Because HDFS was purpose built for Hadoop, it doesn't offer to the Hadoop user the feature set of modern storage platforms that were designed to persist and manage data within the context of multiple, IT production data center use cases. The attributes of data center-grade storage systems could be of great value to Hadoop users in the following ways:

Enhancing data protection and disaster recovery capabilities

HDFS relies on the creation of cloned data copies (usually three) at ingest to recover from disk failures, data loss scenarios and related outages. While this process does allow a cluster to tolerate disk failure and replacement without an outage, it slows data ingest operations, negatively impacts time to information, and still doesn't totally cover data loss scenarios that include data corruption. It also makes for very inefficient use of storage media—a critical concern when users wish to persist data in the cluster for up to seven years as may be required for regulatory compliance reasons.

Modern storage platforms offer automated internal data protection at scale using erasure coding as well as external data protection capabilities (synchronous and asynchronous replication) without the reliance on creating three data copies on ingest. Internal and external automated tiering allows for efficient use of storage resource throughout years of operation.

Making Hadoop clusters more efficient users of compute and storage resources

HDFS binds compute and storage together to minimize the “distance” between processing and data for performance at scale. However, this results in some unintended consequences when HDFS is used as a long term persistent storage environment. To add storage capacity in the form of data nodes, an administrator must add processing and networking resources as well, whether or not they are needed. This tight binding of compute and storage also limits an administrator’s ability to apply automated storage tiering to take advantage of solid state disk at scale.

Integrating a modern storage platform with Hadoop allows users to scale storage independently and as needed without over-provisioning compute and networking resources. Tiered performance via the use of SSD plus HDD within the same system also allows users to balance storage resources and greatly reduce the need to overprovision storage for performance.

More analytics in less time

One of the major advantages of using Hadoop for analytics applications lies in its ability to run queries against very large volumes of unstructured data. For that reason, Hadoop is often positioned as a “Big Data Lake.” The idea is to copy data from active data stores and move the copies to the data lake. However, this process can be time consuming (several hours to days) and network resource-intensive depending of the amount of data.

One way to solve this problem is to consolidate the storage for the multiple applications producing the data with Hadoop storage, eliminating the time-consuming need to create, track, and move data copies over a network. This can be done with a modern storage system that supports industry standard access protocols like NFS, SMB, iSCSI, S3 and Swift. And when integrated with Hadoop, the data created by Hadoop applications is then immediately available to other business applications users. Using a multipurpose storage environment that offers many use cases simultaneously has the further advantage of not requiring modification of the transactional data architecture on which an enterprise may be dependent.

Reducing complexity

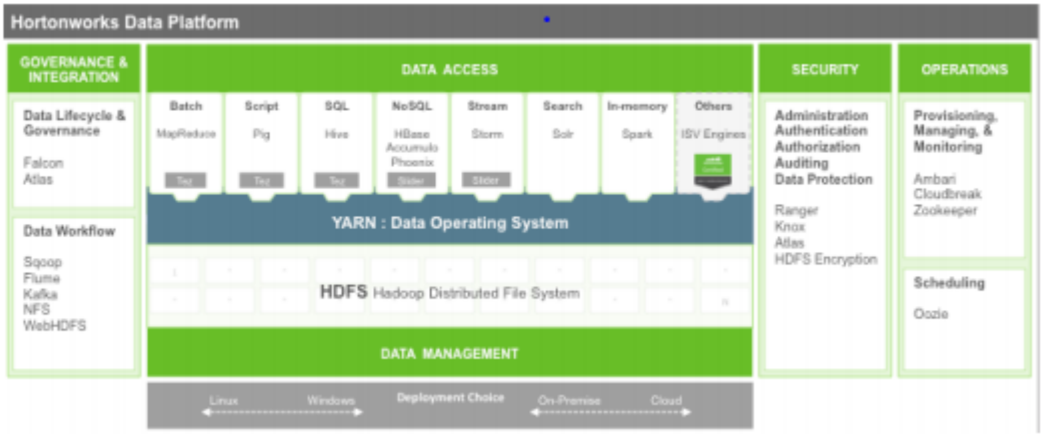
Open source communities create add-on projects to add new functionalities and address deficiencies. In Hadoop, DistCp can be used for periodic synchronization of clusters across WAN distances but requires manual processes to reconcile differences when inconsistencies occur as they will over time. Falcon addresses data lifecycle and management. From the stand point of the Hadoop administrator however, they are typically learned and managed as separate entities. Each has a lifecycle of its own that requires tracking, updating and administering. Enterprise Hadoop administrators will naturally gravitate to simplicity in this regard. Using a storage environment that has already built-in features such as tiering to low cost storage including Cloud, Snapshot integrated data protection, and global data sharing via a common name space, simplifies management and reduces opportunities for error.

These added values are gained through the integration of a modern scale-out, shared storage platform with the Hadoop cluster. Under this scenario, the information sharing and data persistence features of the storage platform are applied to Hadoop without adding, integrating and managing more projects to realize the same outcomes.

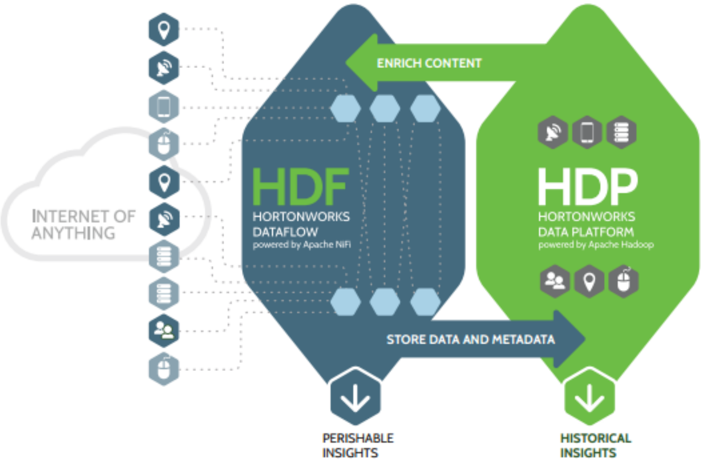
As a case in point, we review the Hortonworks distribution Apache Hadoop, looking specifically at latency-sensitive use cases such as the ones already mentioned. We then examine the newly announced support of IBM's Spectrum Scale storage platform by Hortonworks to address the growing number application scenarios for Hadoop.

Hortonworks HDP and HDF for Today's Business Initiatives

Hortonworks' Apache Hadoop distribution approaches Big Data from the standpoint of differentiating the analysis of data at rest on the one hand and data in motion on the other. The Hortonworks Data Platform (HDP), built on Apache Hadoop, provides a suite of essential Hadoop capabilities that are typically required for enterprise production IT environments. These capabilities include data access and management, security, simplified operations, and adherence to governance practices and mandates (see Figure 2 below). HDP addresses data at rest.



Hortonworks Dataflow (HDF) is based on Apache NiFi technology. NiFi addresses the need to accelerate and simplify the flow of data between systems. HDF is a single combined platform for data acquisition, simple event processing, transport and delivery and is well-suited to tackle and converge the diverse and complex dataflows generated by the growing list of real-time data sources such as smart phones and sensors. When integrated with HDP, HDF captures the often transient data generated by these sources and presents this data to HDP. The combination (see Figure 3 below) results in a single integrated platform for the acquisition, processing, and persistent data storage needed to enable today's business initiates such as Customer 360, IoT, and Industry 4.0. HDF enables the real-time collection and processing of transient data while HDP processes, stores and enriches this data with historical data.



The point of this combination is to accelerate the delivery of information and actionable insights in the context of these new business initiatives as well as the typical Hadoop batch processing-oriented applications. It is therefore a multi-purpose environment that can be enhanced using a scalable, multi-purpose, data center grade storage platform.

To that end, Hortonworks and IBM have announced a support agreement whereby IBM's Spectrum Scale software-defined storage platform can be used to enhance the Hortonworks HDP/HDF for increased user productivity, lower TCO through better storage efficiency, simplified data management and enhanced business continuity.

IBM Spectrum Scale for Hadoop

IBM Spectrum Scale storage software is an enterprise-grade platform for file and object storage and data management. It is based on a parallel file system to provide scaling of performance and capacity in a scale-out architecture. It provides transparent support for Hadoop's HDFS storage layer. When integrated with Hortonworks HDP, its major attributes as a persistent storage environment for production-level Hadoop data repositories include:

Single global name space to support small to large scale Hadoop deployments within a single Spectrum Scale environment by adding new nodes to the cluster.

Unified storage environment—support for both file and object-based data storage. Data access methods include POSIX, NFS, SMB, S3, and Swift

Snapshots at the file system or file set level and backup to an external storage target (backup appliance and/or tape) are also supported.

Synchronous and asynchronous data replication at LAN, MAN and WAN distances with transactional consistency

Automated cloud storage tiering for transparent cloud storage tiering to cloud-based object storage or public cloud storage with automated, policy-driven data movement between storage tiers. Tape is also supported as an additional archival storage tier.

Non-disruptive Operation is supported for active file management and file placement optimization at the global namespace level without requiring a system outage. Rolling upgrades without a system outage are also supported.

Policy-driven data compression (i.e. which files, when and how controlled by the system administrator) can be implemented on a per file basis for an approximately 2x improvement in storage efficiency and reduced processing load on Hadoop cluster nodes.

Storage-based security features include data at rest encryption as an option and secure erase as well as LDAP/AD for authentication. Authentication and authorization via Active Directory or other LDAP source is also supported.

Simple GUI-based management for the storage environment that includes automated resource provisioning and storage system performance monitoring

IBM zSystem integration for users looking to integrate IBM zSystem mainframe data with Hadoop

While Spectrum Scale brings the benefits of enterprise data management to Hadoop clusters, it does not need to completely replace HDFS. It can co-exist with Hadoop clusters deployed on HDFS by presenting a single name space spanning the storage managed by HDFS and Spectrum Scale.

Evaluator Group Assessment:

Hadoop visibility is increasing within the enterprise, driven by its growing applicability to the Digital Enterprise initiatives that can now be found across industry segments. Pressure builds

from business groups that want to do new kinds of data analytics and have heard that their competitors are doing it on Hadoop.

While generally not thought of this way by Hadoop users and administrators, the Hadoop storage environment a critical consideration for these initiatives. We believe that IBM'sSpectrum Scale can add considerable value to the overall Hadoop storage environment in terms of its ability to consolidate data from diferent source and present a coniguous data layer to Hadoop as well as other systems contributing to analytics applications on Hadoop. Data protection and business continuance capabilities enabled by transactionally consistent synchronous and asychronous replication contribute to the perception that Hadoop can indeed stand up to the requirements of enterprise production IT data center deployments. Therefore we see that the partnership that IBM and Hortonworks have formed to certify Hortonworks HDP on IBM Spectrum Scale and will accomplish that end.

Enterprise IT must feel comfortable in managing Hadoop as a platform for production analytics applications. An enterprise data center grade storage system integrated with the HDFS storage environment can address many of their requirements. Here we have shown that IBM's Spectrum Scale softwarecan be used to simplify, protect, and manage Hadoop data resources in ways that are familiar to enterprise IT and conform to their existing data management policies and practices.

About Evaluator Group

*Evaluator Group Inc. is dedicated to helping **IT professionals** and vendors create and implement strategies that make the most of the value of their storage and digital information. Evaluator Group services deliver **in-depth, unbiased analysis** on storage architectures, infrastructures and management for IT professionals. Since 1997 Evaluator Group has provided services for thousands of end users and vendor professionals through product and market evaluations, competitive analysis and **education**. **www.evaluatorgroup.com** Follow us on Twitter @evaluator_group*