SOLUTION BRIEF

Hortonworks Enterprise Data Warehouse Optimization

UPGRADE YOUR ENTERPRISE DATA WAREHOUSE FOR THE DEMANDS OF BIG DATA

The era of Big Data places unprecedented demands on traditional data architecture. Powered by Hortonworks Data Platform (HDP®), the Hortonworks solution for Enterprise Data Warehouse (EDW) optimization lets you overcome challenges of cost, complexity and expansion while extending the value of your existing data warehouse (EDW).

- **Cost**—As organizations seek to capture and use more data, efficiency of storage becomes a key factor in Big Data capacity. Hortonworks makes it possible to reduce hardware and software costs while ensuring that all data can be accessed and analyzed reliably.
- **Complexity**—The integration of new data sources allows you to deliver more powerful analytics and insights for the business—but it also increases data movement and transformation steps. Hortonworks lets you shift data extraction, loading and transformation from your EDW to Apache Hadoop® with schema-on-demand for both known and unknown structures.
- **Expansion**—The structure and scale of the new Big Data domain can make it more difficult to apply existing tools, and more risky to enable agile functionality for traditional systems. Hortonworks provides a mechanism to ingest, store and process any type of structured or unstructured data within an integrated ecosystem.

In this way, Hortonworks allows you to cut the costs associated with your data architecture while extending new analytic value to the enterprise—with the scalability to meet the needs of new data sources, advanced workloads and emerging data science methods.

ADDRESS KEY BIG DATA USE CASES

- **Archive**—Move cold or less frequently used data into Hadoop storage and access it on-demand so you can store more data, longer, while controlling costs.
- **Onboard**—Shift expensive ETL functions from your EDW to Hadoop to leverage compute and operational efficiencies—and free EDW capacity for high-value analytics.
- **Enrich**—Store, process, transform and feed new types of data, from new sources, into your EDW to unlock new analytic value.
- **Interactive Analytics**—Deliver fast, interactive queries, slice and dice analysis on data stored in your Hadoop data stores–accessible from your standard BI tools.

SOLUTION ELEMENTS

- **Hortonworks Data Platform**—Powerful open Hadoop capabilities for data governance and integration, data management, data access, security and operations—architected for deep integration with your existing data center technology.
- **Syncsort**—High-performance ETL software to access and integrate all your enterprise data on HDP.
- **Jethro Data**—High performance analytics engine for interactive BI ON Hadoop to bridge the gap between business users and their data.
- **Professional Services**—Expert guidance and support to maximize the value of the full tested and validated Hortonworks data architecture optimization solution.

UPGRADE YOUR BIG DATA ARCHITECTURE WITH THE POWER OF HADOOP

Hortonworks Data Platform and Hadoop make it possible for your organization to store and analyze data of all types, from any source, with the performance, scalability and cost efficiency your business demands.

Archive Data in Hadoop
EDW costs are often bloated by the storage of cold legacy data with less need for frequent access or real-time performance. In addition, EDW infrastructure may not be able to accommodate non-structured and semi-structured data such as media, log and email files, or data from acquired companies, without extensive modifications to established schema.

Hadoop can play a valuable complementary role to your EDW by serving as low-cost commodity storage for infrequently used data, and providing schema-on-demand for new types of data. Archived data remains secure and governed within HDP, where it can be accessed directly using Tableau business analytics without first being moved to a traditional repository, and blended with other data across HDP and other sources to enable exploration, visualization and insights.

Onboard ETL Processes to Hadoop
Valuable high-performance EDW processing capacity is typically consumed by relatively low-value ETL workloads performed on rarely used data. Hadoop offers a more cost-effective way to accommodate diverse data types without taxing your EDW. Data can be landed initially and stored in Hadoop, with much of the transformation burden deferred until the data is required for analysis, at which point schema can be defined on demand as needed. By shifting ETL from EDW to HDP, you can conserve high-performance EDW processing cycles for high-value analytical workloads.

Enrich theValue of Your EDW
Much of the strategic impact of Big Data is driven by the ability to extract insights from new types of data, from sensor and server log data to clickstream, sentiment analysis and fraud detection data. However, this value can be lost when the EDW is unable to accommodate a data type in its predefined schema.

Hadoop allows your data architecture to ingest and refine virtually any type of data for analysis in Hadoop, the EDW or any other analytical system. This vastly expands the Big Data use cases available to business analysts, including both real-time and historical data, to drive better outcomes today and long into the future.

Fast Analytics on Hadoop
Advances in Hadoop and its ecosystem of partner products now enables users to access data in Hadoop via familiar BI tools like Tableau, Qlik or Cognos. Jethro seamlessly fits on top of your HDP cluster to provide rapid, interactive style slice and dice analytics on all the data. With it's dynamic index and query optimization, the system continuously tunes itself and optimizes query performance without the overhead of manual cube and index maintenance. With Jethro, users can easily query and analyze their big data to create new insights to improve their business outcomes.

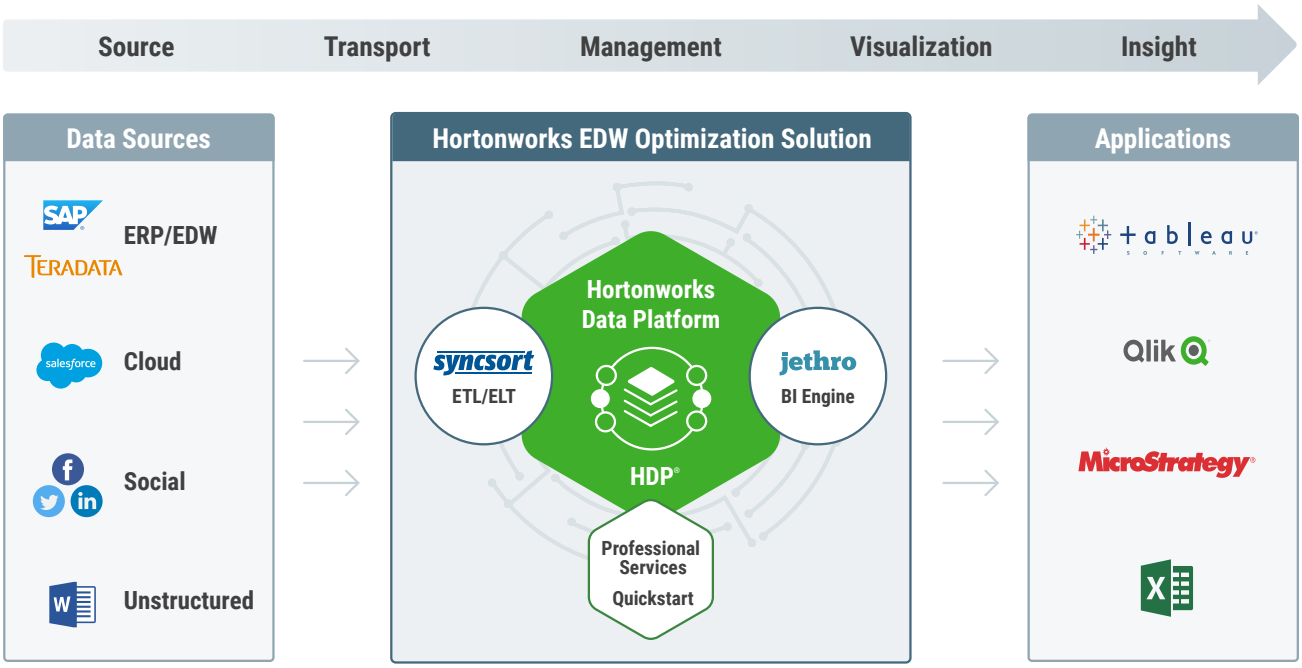

Figure 1: The EDW Optimization solution by Hortonworks, Syncsort and Jethro

VALIDATION

Symantec uses the Hortonworks solution to ingest and process 500,000 security log messages per second (40 billion messages per day) with average time to analysis of two seconds.

The Fuse innovation lab at Cardinal Health has used the Hortonworks solution to optimize its data architecture and enrich its existing data with freely available public datasets.

Neustar uses the Hortonworks solution to capture 100% of its telecommunications network data and retain it for two years—150x more than it could previously store—while saving millions.

COMPLEMENTARY PARTNER TECHNOLOGIES

The Hortonworks solution for data architecture optimization incorporates integrated partner technologies to extend its value for customers. Tested and validated under the Hortonworks Certified Technology Program, these capabilities help form a complete, production-ready solution for Big Data insight.

SYNCSORT DMX-H HIGH-PERFORMANCE ETL SOFTWARE

Designed from the ground up for Hadoop, Syncsort DMX-h simplifies big data integration with everything you need to access and integrate all your enterprise data on Hortonworks Data Platform. Syncsort integrates with Hadoop and HDP directly through YARN, making it easier for users to write and maintain MapReduce and Spark jobs graphically. An intuitive design interface and a Hadoop native runtime engine optimize the performance and efficiency of data integration processes while simplifying their creation and maintenance of these processes.

- **Access**—Ingest data from sources including mainframes, RDBMS, MPP, JSON, Avro/Parquet, NoSQL and more. Import hundreds of tables to HDFS in a single job.
- **Integrate**—Design streaming and batch processes in a single interface with fast multi-data source join and sort capabilities.
- **Comply**—Improve security and governance through seamless integration with Kerberos and Apache Ranger, and by automatically updating HCatalog when loading to Apache Hive™, Apache Avro™ and Apache Parquet.
- **Simplify**—Mask underlying complexity and reduce the need for Big Data integration expertise with design-once, execute-anywhere.

JETHRO DATA

Customers use Jethro to deliver interactive BI Service at attractive Hadoop costs. Jethro transparently supports queries, for thousands of concurrent users, analyzing tens of billions of rows. All that with interactive response times measured in seconds.

- **Interactive Performance**—Jethro's Cost-Based Optimizer combines three different strategies to deliver interactive performance across all types of queries.
  - Full Indexing: every column is automatically indexed.
  - Auto Cubes: every aggregation is automatically turned into a small cube
  - Result Cache: every query result is automatically cached.
- **Seamless Compatibility**—Access your data with popular BI tools like Tableau, Qlik, Microstrategy.
- **High Concurrency**—Supports 1,000's of concurrent BI users running queries with no IT or business data engineering.

About Hortonworks

Hortonworks is an industry leading innovator that creates, distributes and supports enterprise-ready open and Connected Data Platforms and Modern Data Applications that deliver actionable intelligence from all data: data-in-motion and data-at-rest. Hortonworks is focused on driving innovation in open source communities such as Apache Hadoop, Apache NiFi and Apache Spark. Along with its 2,100+ partners, Hortonworks provides the expertise, training and services that allow customers to unlock transformational value for their organizations across any line of business.

Contact

For further information visit
hortonworks.com

+1 408 675-0983
+1 855 8-HORTON
INTL: +44 (0) 20 3826 1405

HORTONWORKS

02/2018