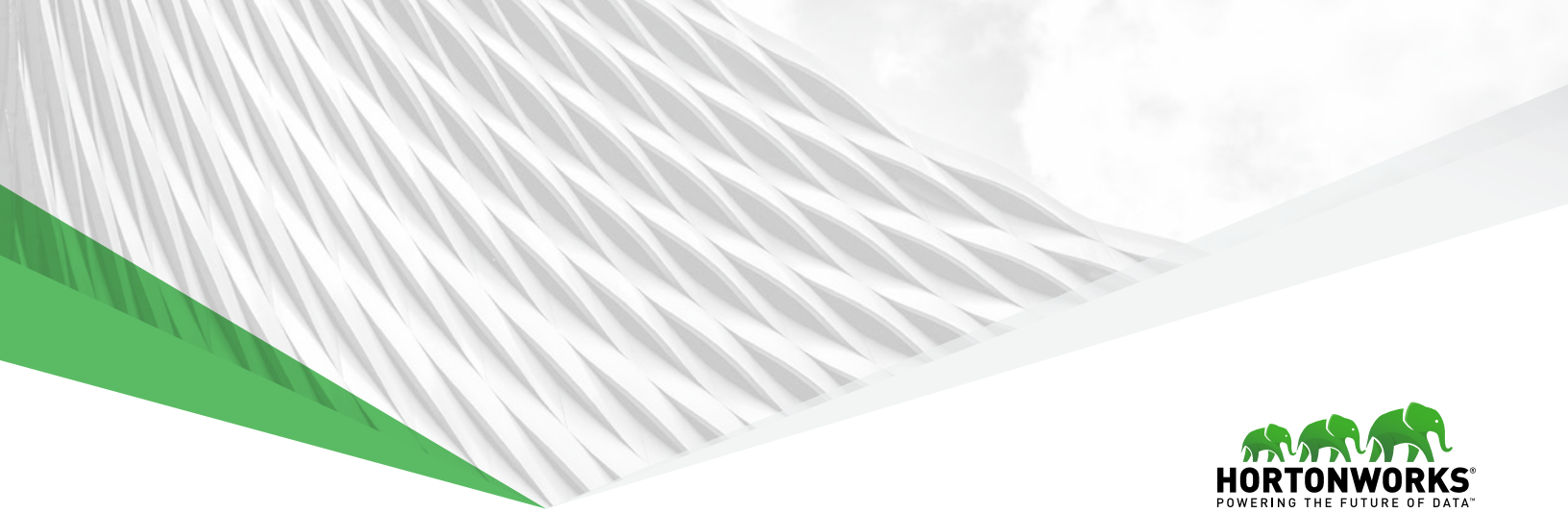

DATA ARCHITECTURE OPTIMIZATION

A HORTONWORKS WHITE PAPER
MARCH 2016

Contents

DATA ARCHITECTURE OPTIMIZATION

DATA ARCHITECTURES WILL EVOLVE TO MAKE POSSIBLE NEW BUSINESS OUTCOMES

The storylines are well-worn: data as an underutilized renewable resource; data transforming industries; every business is a data business. But organizations must deal with the realities of existing data architectures and budget constraints. Are there practical steps to take in order to turn these data storylines into real business outcomes?

Data Architecture Optimization describes an approach to apply Apache Hadoop and related community-driven open source technologies to make data architectures capable of realizing new and improved business outcomes while driving significant cost out of the IT budget.

The rapid growth in data volumes from a wide range of new sources indeed offers disruptive opportunity to those who can put it to use. There is a change in mindset among IT organizations and data architects, who now look to capture all data, keep it longer, and prepare to use the data in new ways as business conditions evolve. These changes create dramatic pressure on traditional data architectures, which were built to support structured data with modest growth.

Hortonworks Data Platform (HDP), built on Apache Hadoop, offers the ability to capture all structured and emerging types of data, keep it longer, and apply traditional and new analytic engines to drive business value, all in an economically feasible fashion. In particular, organizations are breathing new life into enterprise data warehouse (EDW)-centric data architectures by integrating HDP to take advantage of its capabilities and economics.

Optimize your data architecture with Hortonworks Data Platform and Apache™ Hadoop®

The Enterprise Data Warehouse (EDW) became a standard component in corporate data architectures because it provides valuable business insights and powerful decision analytics for front-line workers, executives, business analysts, data scientists, and software developers. For years, the EDW has been the core foundation on which analytics are built within the enterprise. EDW solutions are mature and extremely effective for reporting and data analytics—especially for known transactional data. As new data types and new analytic tools have emerged, however, a new approach for a broader data and analytics architecture in the enterprise has become necessary. With the burgeoning volume and variety of these new data sources and types, the center of gravity for modern data architectures is shifting. This is causing the need not only for new agile tools, but also for an integrated ecosystem solution.

The need for a new approach is evident in the challenges IT executives and business executives face:

- Cost: As data volumes and varieties grow, so do the costs. Often organizations struggle to validate the cost of storing data against the value provided. As the scale of data to be managed outpaces Moore's Law, new architectural approaches are required to keep costs contained.
- Complexity: Data architectures have an expanded role and footprint, integrating many data sources to drive analytics engines. Data movement and transformation steps have multiplied as a result. The world has moved from traditional 'schema on write' for known and fixed data structures to 'schema on demand' for known as well as unknown structures.
- Expansion: New (and often big) data sources have emerged, presenting both structural and scale challenges to the EDW-centric architecture. It can be very difficult to apply existing tools in the new domain, and even more risky to enable agile functionality inside of traditional business-critical systems.

Within any data architecture, the value of the EDW is clear. Businesses require analytics-driven insights and real-time dashboards to run efficiently. Predictive analytics enable successful companies to compete and win. All the while, driven by new competitive pressures, new data sources, and budget constraints, leading organizations are augmenting traditional data architectures with Apache Hadoop as a way to extend capabilities with new sets of data and analytic applications while containing or reducing costs. As an answer to the challenges of cost, complexity, and expansion, organizations are turning to Apache Hadoop to modernize their data architectures and at the same time enhance the value of their existing EDW implementations.

A strained data architecture

In traditional data center architectures, the EDW is a key component in the analytics chain, ingesting data from systems of record, processing it and then feeding key analytic capabilities across data marts, apps and dashboards. This generally accepted approach has served the enterprise well for years. As the landscape of data sources and systems has grown, however, several limitations in this approach have become apparent:

- Cold or rarely used data is stored in expensive, high performance data systems across the architecture. A typical EDW, for instance, dedicates a majority of its storage capacity to cold data. Many organizations face difficult decisions over which data to store and how long to keep data in order to manage costs, while sacrificing the value of deeper analytics enabled by the additional data.
- Relatively low-value Extract-Transform-Load (ETL) workloads are performed on rarely used data and consume significant processing cycles in high performance systems. In a typical EDW, ETL processing can account for over half of processing capacity.
- New types of data, such as clickstream, sensor, and server log data, that do not fit predefined schema in the data architecture and EDW are poorly managed. Only a subset or aggregate of the data is maintained, meaning valuable insights are discarded.

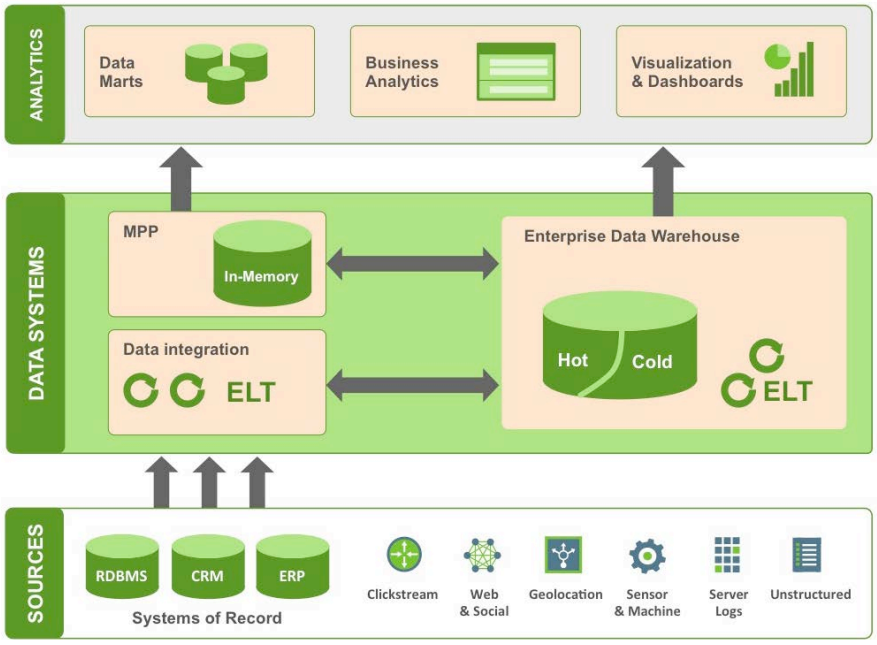

Figure 1: A traditional data architecture

Archive, Onboard, and Enrich

Leveraging Apache Hadoop to build a modern data architecture allows organizations to take practical steps to support the new demands of rapidly growing volumes and varieties of data. These steps include creating an active archive, onboarding ETL processing, and enriching the data architecture with new data sources.

Creating an active archive in Apache Hadoop accomplishes several goals. First, it provides economical storage for cold and rarely used data across the data architecture, freeing up capacity in expensive systems. Second, it allows an organization to capture any type of data and keep it much longer, taking advantage of Apache Hadoop's flexible design allowing storage capacity to be added incrementally as needed. And third, it brings archived data back to life, making it easily accessible.

Onboarding ETL processing to Apache Hadoop offers a similar range of benefits. As with moving data, moving processing from expensive systems frees up capacity in those systems for other tasks, such as high value analytical workloads. Additionally, because Apache Hadoop is a schema-on-demand system, organizations can gain efficiencies by implementing ELT processes, landing all data in Apache Hadoop and deferring much of the transformation burden until the data is required for analysis.

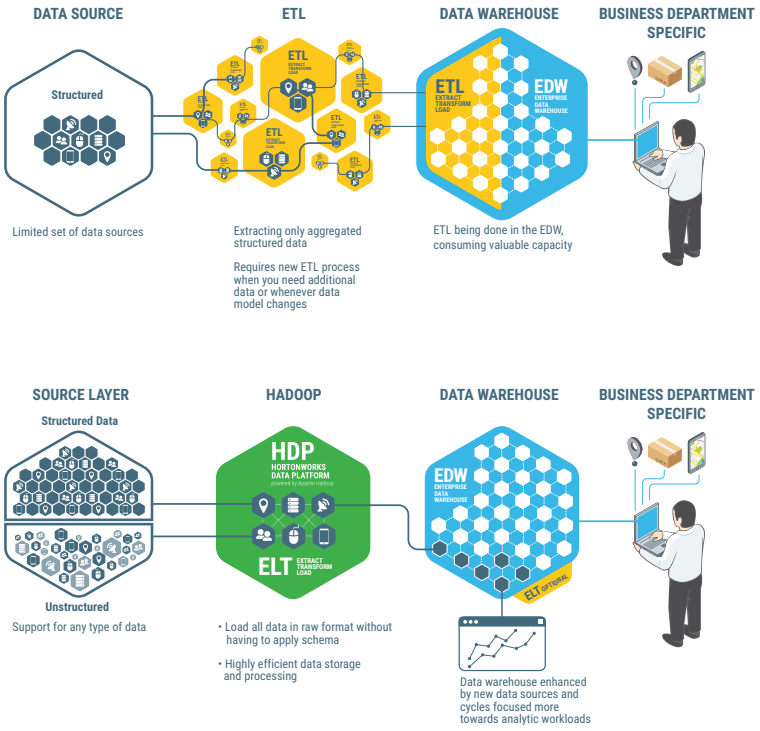

Figure 2: ETL processing before and after Apache Hadoop

Finally, Apache Hadoop allows organizations to enrich the broader data architecture by ingesting new types of data and refining that data to be analyzed in Apache Hadoop, in the EDW, or in any analytical system that can add value to the business. The ability for Apache Hadoop to ingest, store, and process any type of data makes possible the variety of use cases that cause people to declare data is transforming business. And because all of the data and analytical insights are retained, businesses can build on them to drive better outcomes in the long term.

ARCHIVE DATA IN APACHE HADOOP

Move cold or rarely used data to Apache Hadoop as an active archive to reduce costs while expanding the amount of history that can be maintained.

ONBOARD ETL PROCESSES TO APACHE HADOOP

Implement ETL processes in Apache Hadoop, perform more efficient ELT, and reduce costly data movement.

ENRICH THE VALUE OF YOUR EDW

Use Apache Hadoop to refine new data sources, such as web and machine data, so they can be used to fuel your business and expand opportunity.

A data architecture optimized with Apache Hadoop

With Apache Hadoop, the data architect can cut costs associated with the data architecture while extending new analytic value to the enterprise. Apache Hadoop provides linear scale storage and compute so that it can scale to meet the needs of not just new data sources but more advanced workloads and data science methods that have emerged. The benefits of this new architecture include:

1. Move rarely used data to Apache Hadoop and access it on demand, saving on overall storage costs
2. Store more data longer to enhance analytics with deeper information providing better results
3. Store and process new data sources and feed transformed data into your EDW to augment or create wholly new analytic value
4. Onboard ETL processes in Apache Hadoop in order to take advantage of compute and operational efficiencies

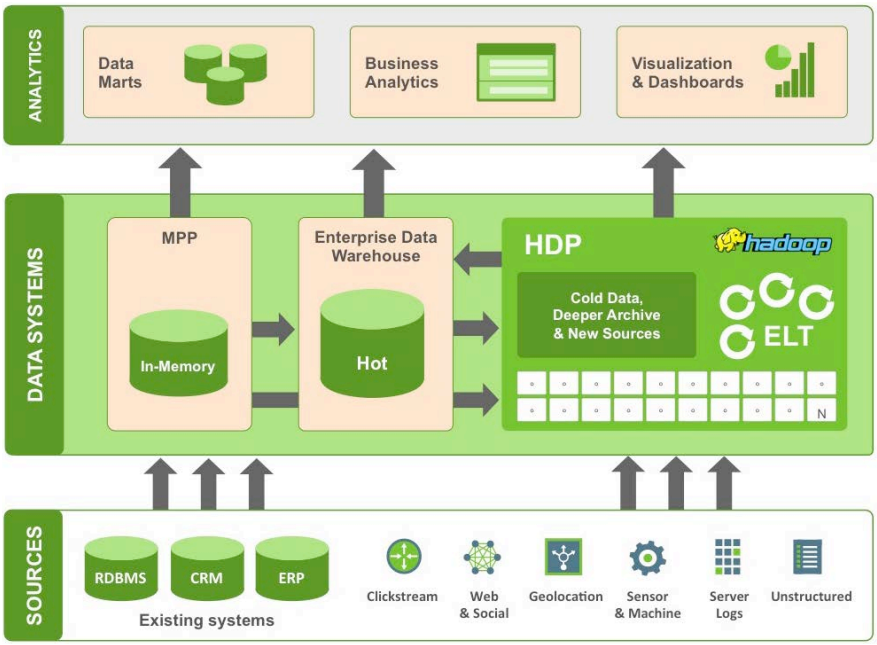

Figure 3: A modern data architecture with HDP

Hortonworks Data Platform and Apache Hadoop

Evolving the data architecture with Hortonworks Data Platform and Apache Hadoop enables your organization to store and analyze data at massive scale, extract critical business insights from all types of data from any source, improve your competitive position in the market, and maximize customer loyalty and revenues.

HDP provides the full suite of essential Apache Hadoop capabilities that are required by the enterprise and that serve as the functional definition of any data platform technology. HDP is architected to integrate deeply with existing data center technology, and includes a comprehensive set of capabilities including Data Management, Data Access, Data Governance and Integration, Security, and Operations. And HDP is developed entirely in the open community, allowing you to take advantage of rapid community innovation and deep integration across the ecosystem, while avoiding proprietary lock-in.

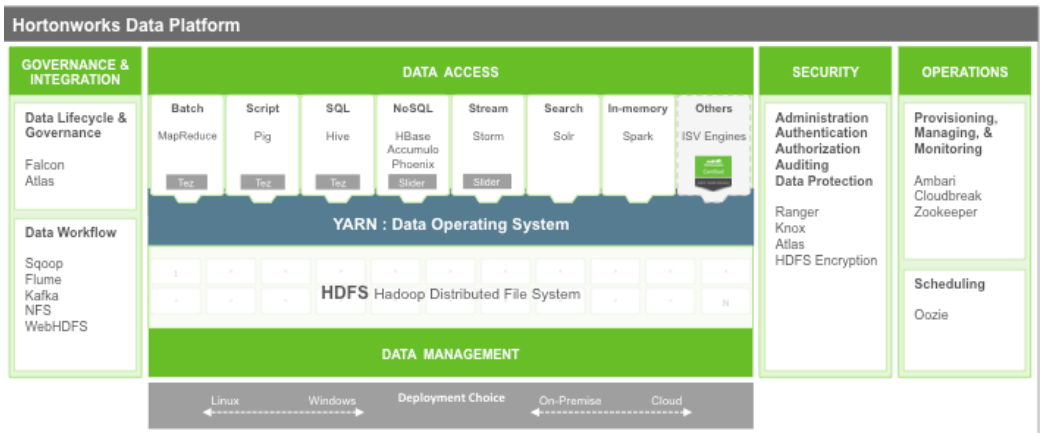

Figure 4: Hortonworks Data Platform capabilities—data governance and integration, data management, data access, security, and operations

About Hortonworks

Hortonworks powers the future of data. Our team develops, distributes and supports Hortonworks Data Platform (powered by Open Enterprise Apache Hadoop) and Hortonworks DataFlow (powered by Apache NiFi). Working together, these connected data platforms provide actionable intelligence through modern data applications. Our team comprises the largest contingent of builders and architects within the Apache Hadoop ecosystem and leads the broader enterprise requirements within these communities. Our connected data platforms integrate deeply with existing IT investments, and Hortonworks co-engineers our technology with other leading data center vendors so that our customers can unlock the broadest set of Big Data opportunities.

Contact

For further information visit
http://hortonworks.com/hadoop/

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

HORTONWORKS
POWERING THE FUTURE OF DATA