BIG DATA: FIVE TACTICS TO MODERNIZE YOUR DATA WAREHOUSE

Current technology for Big Data allows organizations to dramatically improve return on investment (ROI) from their existing data warehouse environment. Today, we have new data storage and management architectures for Big Data designed to meet the challenging analytical needs of the modern enterprise. These new architectures and technologies are capable of managing massive data sets - integrating structured and unstructured data - to deliver real-time capabilities and predictive analytics. Imagine the ability to quickly uncover customer, product and operational insights buried in all of those transactional, social, mobile, and sensor data sources.

Traditional data warehouses were built with online transaction processing (OLTP)-centric technologies and architectures that are 15-20 years old. These data warehouses were never designed to handle the volume, variety and velocity of today's data-centric applications. Over the years, more and more data has been bolted on to these data warehouses, while the query load driven by business intelligence products has increased exponentially. Consequently, this has resulted in brittle, over-burdened, and costly data warehouses that require 6 to 9+ months to add the next data source.

Companies can greatly benefit from new technologies, products, and approaches to modernizing these outmoded, inflexible data warehouses, making them much more responsive to their marketplace. This paper describes five tactics for organizations to begin to modernize their data warehouse operations. This modernization can rapidly lower CAPex and OPex by decreasing data acquisition, maintenance and administrative costs, while improving overall performance, agility and scalability.

REDEFINE

EMC PERSPECTIVE

EMC²

Tactic #1: Embrace the Data Lake

Nothing will have as big a positive impact on your long-term data storage, management and analysis capabilities as Hadoop and the Hadoop Distributed File System (HDFS). Without a doubt, Hadoop is a game-changer from both an IT and a business perspective. For many organizations, the introduction of Hadoop/HDFS into the organization begins with the establishment of the Data Lake. A Data Lake is a storage repository that holds a vast amount of raw data in its native (as-is) format until it is needed.

DATA LAKE APPLICATIONS

Allows integration of unstructured claims descriptions to reduce fraudulent claims

Leverages mobile data to create real-time promotional opportunities

Leverages sensor readings to predict maintenance needs and pre-empt costly downtime

Hadoop/HDFS is a cost-effective, scale-out storage system with natively parallel computing and analytical capability. Traditionally built on commodity clusters, Hadoop/HDFS simplifies the acquisition and storage of diverse data sources, whether structured, semi-structured (e.g., web logs, sensor feeds), or unstructured (e.g., social media, image, video, audio). Figure 1 depicts Hadoop within a modern data architecture.

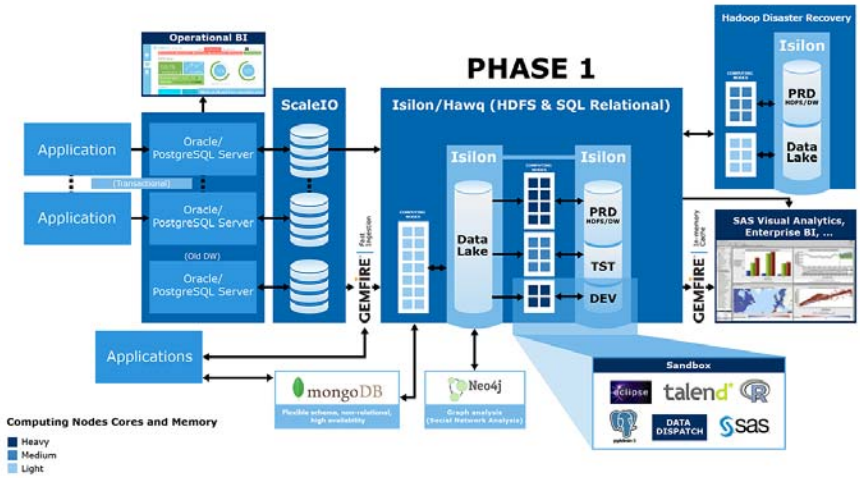

Figure 1: Reference Architecture for the Modern Data Architecture

Hadoop provides the foundation for a modern data architecture. At a high-level, this modern data architecture is comprised of three critical components (see Figure 2):

1. **BI/DW Environment -** this is your traditional data warehouse, which supports the operational and management reporting with respect to understanding "what happened?" types of business questions. This is a production environment with very predictable loads that is SLA-driven and heavily governed. Most organizations look to enforce data transformation, databases, and business intelligence (BI) tool standards at this level in order to drive down costs and ensure an SLA-compliant environment.
2. **Analytics/Sandbox Environment –** this is where your data science team provisions compute environments and desired data sources in order to uncover new customer, product and operational insights. This environment is almost the polar opposite of the BI/EDW environment. It is an exploratory environment with very unpredictable load and usage patterns. It is an environment where the data science team needs to be free to experiment with new data sources, new data transformations and enrichment algorithms, and new analytic models in order to uncover new insights buried in the data and build predictive models of an

   Organization's key business process. It is loosely governed and typically allows the data scientists to use whichever tools they prefer in their exploration, analysis, and analytic modeling.
3. **Hadoop Data Lake –** this is the central repository for all the organization's data (absent the burden of predefining your data schemas). The Data Lake can feed both the production BI/DW environment and the exploratory analytics sandbox as necessary. The immediate modernization opportunity for many organizations is off-loading the ETL (extract, transform, and load) routines from the expensive data warehouse to the Data Lake. Existing ETL routines can be dramatically accelerated using the native parallel nature of Hadoop. And new "data enrichment" processes can

be developed unlocking new metrics (e.g., frequency, recency, sequencing, etc.) that may be better predictors of business performance.

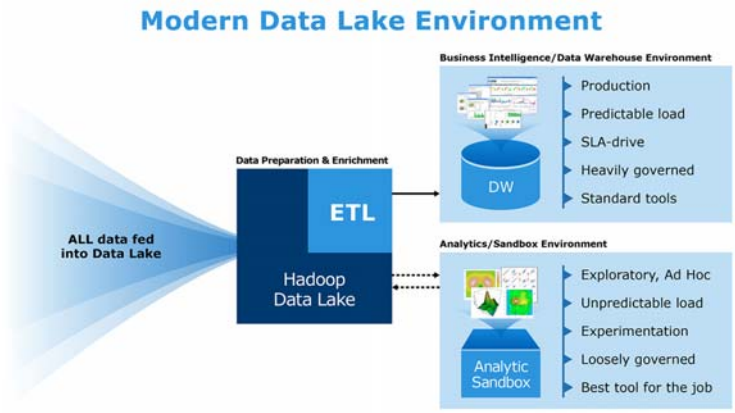

Figure 2: Hadoop as the Foundation of your Data Management and Analysis Architecture

MPP APPLICATIONS

Affords seasonality to forecast retail sales and energy consumption

Allows localization to pin point lending or fraud exposure

Offers Hyper-dimensionality for digital media attribution or health care treatment analysis

Reduces data warehouse administrative overhead and costs, while improving data warehouse agility and performance

Tactic #2: Super-Charge Your Data Warehouse Through Massively Parallel Processing

Many traditional data warehouses are built on OLTP-centric relational data base management systems (RDBMS). Those RDBMS were designed for OLTP data entry environments that operate on a single record at a time (e.g., add, update or delete). Data warehousing is the mirror opposite, requiring access to a massive number of records in order to perform even simple analytics such as trending and comparison examination (current period versus previous period).

In order to make these OLTP-based RDBMS support data warehouse requirements for a massive number of records/transactions, RDBMS vendors resorted to design tricks including materialized views, aggregate tables and indices. The first problem is that the amount of storage, processing power and human resources required to maintain this approach soon dwarves the effort required to load and manage the base data warehouse. The second problem is that a key objective to use "trickle feeds" to create a "real-time" data warehouse is never realized because each time new data is added, the materialized views, aggregate tables and indices for the data warehouse have to be rebuilt. This approach does not provide a real-time environment!

Utilizing massively parallel processing for the data warehouse enables more granular data query, reporting, and dashboard drill-down and drill-across exploration. It allows analysis to be performed on detailed data instead of just data aggregates. The benefits related to agility and performance gains resulting from a traditional OLTP-based data warehouse to a MPP based data warehouse are striking (see Figure 3).

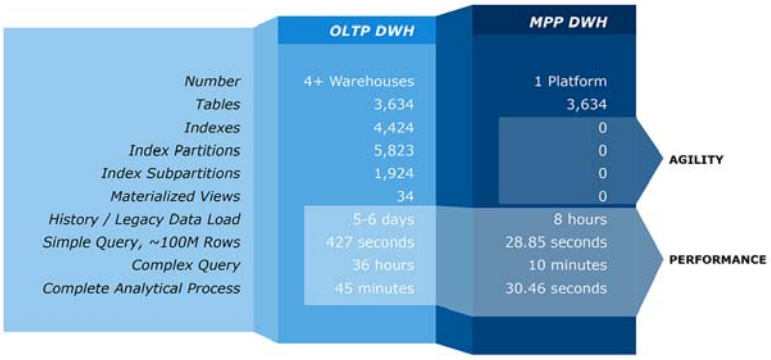

Figure 3: Agility and Performance Results for OLTP DWH versus MPP DWH

Recent developments (e.g., Pivotal HAWQ) now permit organizations to build their data warehouse directly on HDFSs. This approach allows these organizations to benefit from the cost efficiencies, scale-out architecture and native parallelism provided by HDFS, while providing access to the HDFS-based data warehouse using the organization's standard SQL-based BI tools.

On the analytics side, once a model has been developed and business insights have been gleaned from these data sets, simply migrate the analytic model and/or the analytic insights into the existing data warehouse for integration into the current business intelligence environment. Alternatively, the analytic modeling can also be executed on the MPP platform, making it part of the production process.

Tactic #3: Capitalize on In-Database Analytics

A leading development for big data is the advent of in-database analytics. In-database analytics allows the analytic models and algorithms to be executed directly within the database. Doing so eliminates the time and effort required to transform data and move it back and forth between a database and a separate analytics environment.

IN-DATABASE APPLICATIONS

Drives real-time customer acquisition, predictive maintenance, or network optimization decisions

Enables new location-based services

Leverages the Internet of Things to create new products and services

Updates analytic models on-demand, based upon current market or local weather conditions

In-database analytics addresses one of biggest shortcomings when performing advanced analytics – the requirement to move large amounts of data. This data movement issue has forced many organizations and data scientists to work with aggregate tables. This data transfer issue is debilitating to the rapid, "fail fast" analytic exploration, discovery, and model development process. In-database analytics reverses the process by moving the analytic algorithms to where the data is stored. This accelerates the model development, refinement and deployment process. The elimination of data movement delivers substantial benefits that include:

- Reducing the time it takes to move terabytes of data from hours to zero.
- Reducing the processing time required to analyze terabytes of data by a factor of 16. Figure 4 shows an example of going from 193 minutes to only 12 minutes using a 5-processor system.

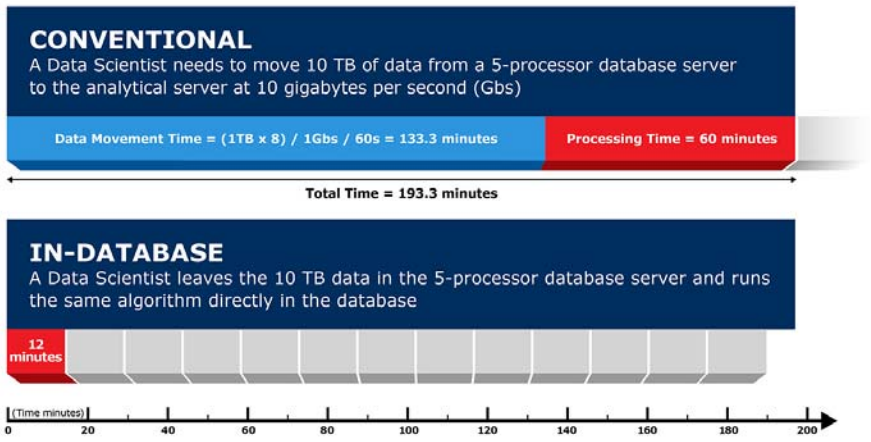

Figure 4: In-Database Analytics Dramatically Speeds Processing Time

In one use case, we transformed several sequential algorithms to a MapReduce job. The execution time dropped from more than 13 hours to 12 minutes, running on 160 cores simultaneously. This is the kind of benefit that in-database analytics can provide.

Tactic #4: Add Unstructured Metrics to the Existing Data Warehouse

An easy way to start building experience with Hadoop is to create new metrics from an unstructured data source that can be fed into the existing data warehouse. This provides the ability to leverage data such as social, mobile, consumer comments, e-mail, doctors' notes, or claims descriptions to identify new metrics that may be better predictors of behavior.

It also simplifies an organization's ability to massage and parse unstructured data (e.g., log files, text files, research publications, physician notes, etc.), to uncover new predictive measures in the unstructured data, and feed that data into the existing data warehouse. These new metrics can then be easily integrated into an organization's existing business intelligence queries, reports, dashboards, and analyses (see Figure 5).

UNSTRUCTURED METRICS APPLICATIONS

Leverages customer interests, passions, associations, and affiliations to improve acquisition, maturation, retention, and advocacy development

Integrates sensor-generated data into your manufacturing, supply chain, or predictive maintenance models

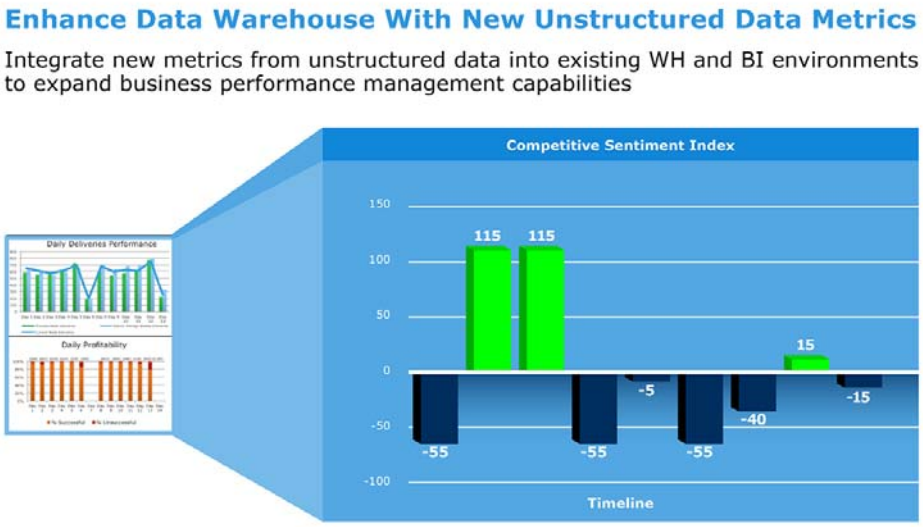

Figure 5: Example of Integrating Social Media Metrics into the Existing BI Environment

Tactic #5: Use Data Federation to Extend Your Data Warehouse

Continued advances in the area of data federation are allowing organizations to extend the data warehouse view to permit access to external data sources on an as-needed basis. This "federated data warehouse" gives an organization quick access to seldom used data sources without going through the process of moving that data into the existing data warehouse or the Data Lake. This approach allows data outside the existing data warehouse to be accessed and analyzed.

For example, you may not want to download all of your detailed social media data from sites such as Facebook, Twitter, Pinterest, and LinkedIn into your Data Lake. Instead, organizations can simply establish a conduit (via the social media site APIs) to those sites for gaining access to relevant detailed data as needed (see Figure 6).

DATA FEDERATION APPLICATIONS

Provides access to infrequently requested, massively large data sources (web, social)

Supports one-off business analytic requests

Allows test and validation of business use cases prior to moving into the data warehouse

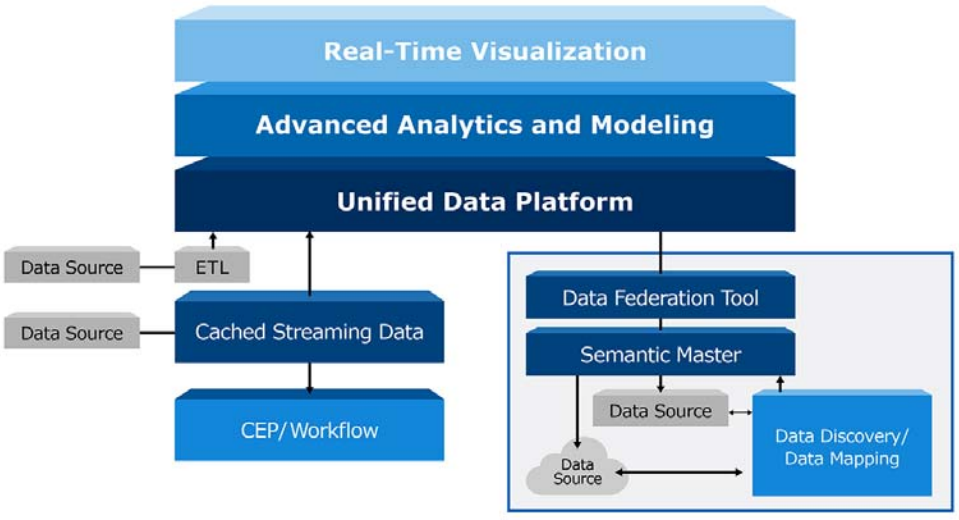

Figure 6: Use Data Federation for on-demand access to seldom-used data outside your DW

Conclusion

Traditional data warehouses are unable to meet the growing need of the modern enterprise necessary to integrate and analyze a wide variety of data being generated from social, mobile and sensor sources. More importantly, these data warehouses struggle to answer the forward-looking, predictive questions necessary to run the business at the required levels of granularity or in a timely manner to remain competitive. This paper highlighted five simple ways for organizations to begin to benefit from the advantages of a modernized data warehouse architecture (see Figure 7).

Each of the five tactics described within this paper are independent of one another and each delivers its own business benefits. Organizations who employ these tactics should see improved CAPex and OPex costs through decreasing data acquisition, maintenance and administrative costs, while improving overall performance, agility and scalability.

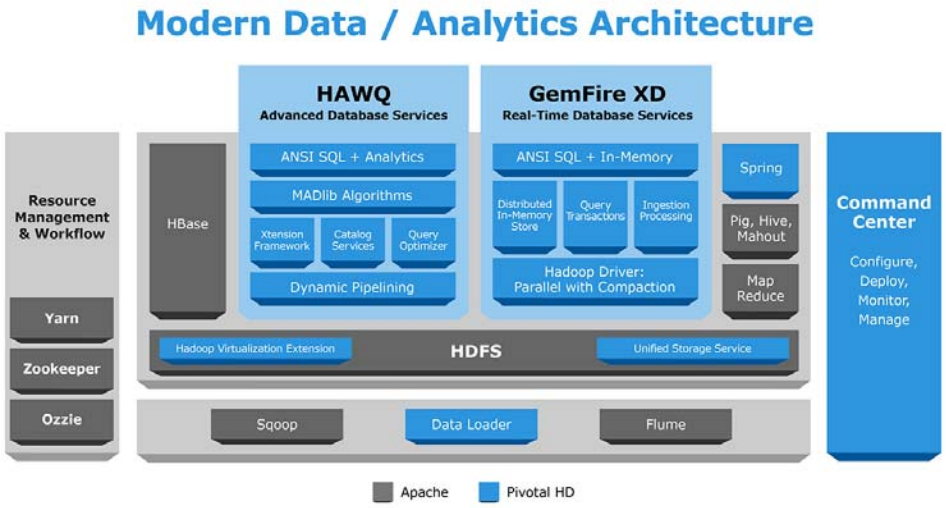

Figure 7: Parse Unstructured Data to Create New Metrics for Your Data Warehouse

About EMC Global Services

EMC Global Services accelerates the software-defined enterprise through world-class technical expertise and service capabilities that deliver well-run hybrid clouds, big data solutions, empower ITaaS providers, and enable new digital-era applications. Our 16,000+ services experts worldwide, plus global network of partners, have the skills, knowledge, and experience organizations need to get the maximum value from their EMC technology investments—with an unending commitment to an exceptional total customer experience through service excellence.

CONTACT US

To learn more about how EMC products, services, and solutions can help solve your business and IT challenges, contact your local representative or authorized reseller—or visit us at www.emc.com.

Published in the USA. 10/2014 EMC Perspective H13622