WHITEPAPER

BIG DATA

PERSISTENT

How to Enhance Traditional BI Architecture to Leverage Big Data

Contents

Executive Summary..... 1

Traditional BI - DataStack 2.0 Architecture..... 2

Benefits of Traditional BI - DataStack 2.0..... 2

Shortcomings of Traditional BI - Datastack 2.0..... 3

Benefits of Big Data..... 4

Big Data - BI Integration Challenges..... 5

Enhancing BI Architecture to Harness the Power of Big Data..... 5

Benefits of Enhanced BI Architecture..... 7

Conclusion..... 7

Executive Summary

Business Intelligence (BI) has become an integral part of enterprises as it catalyzes effective business decisions. Enterprise systems have standardized technologies to store transactional data on data warehouses and make the data available for various types of analysis using BI tools. Although industry-standard Business Intelligence architecture has been around for more than a decade, it needs to be revisited with the advent of Big Data. Big Data provides a cost effective and highly scalable platform to analyze all data formats; and its close integration with BI systems is a big boost to traditional architecture.

This paper discusses how to leverage powerful Big Data platforms on a traditional BI system architecture. In addition, this paper explains how making incremental changes to conventional BI architecture as a part of a Big Data rollout strategy can benefit your ROI. This paper will help system architects, program managers, and CIOs make informed decisions regarding implementing Big Data technologies in their existing business applications and enterprise architecture. Although, the paper focuses on enterprise environments, the recommended architecture can be effectively implemented by ISVs for BI products.

Introduction

Business intelligence has become an integral part of enterprises to help businesses make effective decisions. Although industry standard Business Intelligence architecture has been around since more than a decade, with numerous software vendors and IT companies providing cost effective and efficient BI solutions to enterprises, it now needs to be revisited with the advent of Big Data.

Although industry standard Business Intelligence architecture has been around since more than a decade, with numerous software vendors and IT companies providing cost effective and efficient BI solutions to enterprises, it now needs to be revisited with the advent of Big Data.

Over the past few decades there has been a gradual evolution of data management technologies from OLTP (Online Transaction Processing) systems to data warehousing and BI, with the latest trend being Big Data. Big Data has opened a new paradigm for storing and analyzing high volumes of data. Almost every big enterprise is now experimenting and discovering use cases with Big Data in its own business domain and data processing environment. Big Data platforms are commonly used in supporting mainstream business for internet based companies that need to process extremely high volumes of data extending up to hundreds of petabytes. However, in enterprises, Big Data is still being used to solve specific data processing and storage problems, rather than being integrated with the enterprise's data architecture. As a whole, Big Data platforms for enterprises have significant benefits and applications for mainstream data processing.

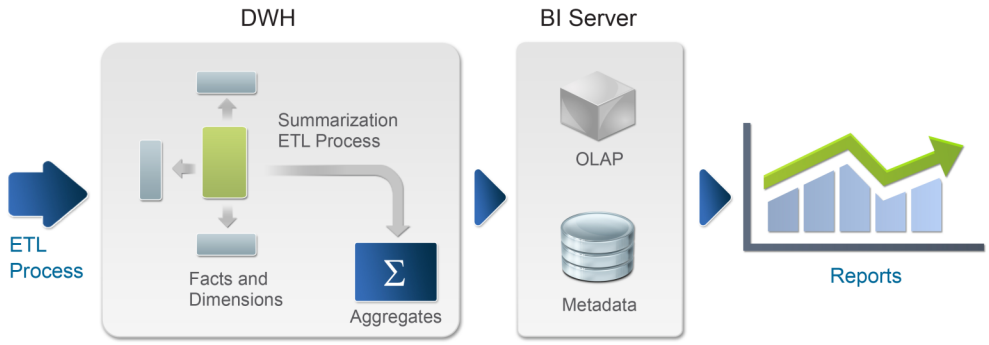
The purpose of this paper is to help system architects, program managers and CIOs take advantage of Big Data technologies in existing business applications. In this paper, we describe architecture nuances and introduce various technical components to approach such scenarios.

Traditional BI - DataStack 2.0 Architecture

We define DataStack 2.0 as the industry standard of BI architecture, built on relational databases that hosts data warehouses and reports generated by BI tools. The figure below depicts traditional BI architecture and the various components of DataStack 2.0 are listed below:

DataStack 2.0 - An industry standard of BI architecture, built on relational databases that hosts data warehouses and reports generated by BI tools.

Figure 1: Traditional BI Architecture



- **Data warehouse (DWH)**
  DWH hosts transactional data in a dimensional form that is easily queried by metadata driven and self-service reporting tools. Aggregated data is also maintained in the warehouse to optimize standard reports and ad-hoc analysis.

- **ETL (Extract-Transform-Load) process**
  The ETL process is responsible for keeping the data warehouse in sync with operational systems. The process extracts changed data from OLTP systems, cleanses it, relates data from various sources, transforms the inbound data to warehouse schema format, and loads it to the data warehouse.

- **BI tool**
  BI tool exposes DWH data to business users in the form of user friendly reports and analytics. OLAP cubes and metadata are supported to provide ad-hoc analysis capabilities to end users. Analytics tools can run on the warehouse to generate trends, forecasts, and discover patterns.

Benefits of Traditional BI - DataStack 2.0

DataStack 2.0 has stood the test of time in the BI and OLTP space due to various reasons including:

- **Strong Support for Standardized Interfaces** - Interfaces such as SQL and OLAP, are intuitive, powerful and supported by standard BI and analytics tools.
- **High Data Consistency** - Various transaction management features enable relational databases to consistently host data in exceptional scenarios, hence providing high-data consistency and confidence to the business on reporting computational data.
- **Standardized Tools** - Maturity and ecosystem of products that support DataStack 2.0 are overwhelming. Open source and commercial data management tools for Data Modelling, ETL, BI, MDM (Master Data Management), Metadata Management, Data Quality, Data Migration, and Tuning are extensively used as industry practice.
- **Improved Performance** - Recent developments in parallel databases such as shared-nothing architecture, column-oriented storage, and in-memory support have improved performance.

Shortcomings of Traditional BI - Datastack 2.0

Although there are significant benefits to traditional BI, the strong market penetration of DataStack 2.0 is a barrier for Big Data adoption. As a result, there are two distinct set of use cases in the market - traditional BI and Big Data. While major market players in ISV space are offering Big Data solutions, close integration between their BI products and Big Data is not supported, resulting in enterprises seeing these as two unrelated categories of data stores.

While major market players in ISV space are offering Big Data solutions, close integration between their BI products and Big Data is not supported, resulting in enterprises seeing these as two unrelated categories of data stores.

Despite numerous strong factors supporting DataStack 2.0 in the market, there are a few major concerns:

- **Scalability & Performance** - Constant investments in maintaining performance and scalability remain a challenge. Practical scenarios such as avoiding degradation of report responses with increase in data volume can't be addressed easily.
- **Additional Cost** - DWH scale-up leads to substantial hardware and software license costs. Tool license fees and maintenance costs, to manage large volumes of data, pose additional challenges. Enterprises have to allocate sizable funds for tuning and hardware upgrades.

- **Analytics Platform** - Lack of support for consolidated analytics platform. Although analytics tools can be run on data warehouses, they require independent powerful data crunching servers, which come with additional hardware, software costs and maintenance.
- **Custom Analytics** - Performing custom analytics on DWH it not easy, and brings down performance of the DWH.
- **Archived Data Management** - Overhead exist in managing historical data for compliance and reporting purpose.

As Big Data can mitigate many of these issues, the shortcomings of traditional BI system can be overcome by tight integration between the two systems.

Benefits of Big Data

Big Data adoption is quickly gaining momentum for data intensive applications. Simply put-if you are willing to compromise on certain relational database features such as support for complete SQL syntax and strict consistency levels, you can save on the licensing cost by investing a little in managing Big Data servers. Here are some of the multifold benefits of Big Data:

If you are willing to compromise on certain relational database features such as support for complete SQL syntax and strict consistency levels, you can save on the licensing cost by investing a little in managing Big Data servers.

- **Scalable up to Hundreds of Petabytes**
  DataStack 2.0 architecture cannot support such high volume data events with advance parallel database architecture. A proof of concept can be developed on Big Data with a single node and can be scaled easily.
- **Cost Savings**
  With most of the Big Data platforms available under open source licenses, cost of ownership is significantly less than that of traditional BI architecture. Many enterprises are adopting Big Data due to low cost, flexibility of configuration and application development by cutting down dependencies on software vendors.
- **MapReduce Programmatic Interface**
  MapReduce provides powerful programmatic interfaces for custom data processing, which goes beyond SQL capabilities. This is especially helpful for analytics use cases. Big Data platforms support powerful and simple interfaces for data querying using JSON and trimmed down version of SQL to facilitate quick development of applications and analytics solutions.
- **Support for Semi-Structured and Unstructured Data**
  Flat files and unstructured data form a big chunk of information in enterprises. Big Data can integrate these missing pieces to support complete enterprise wide analytics.

Big Data - BI Integration Challenges

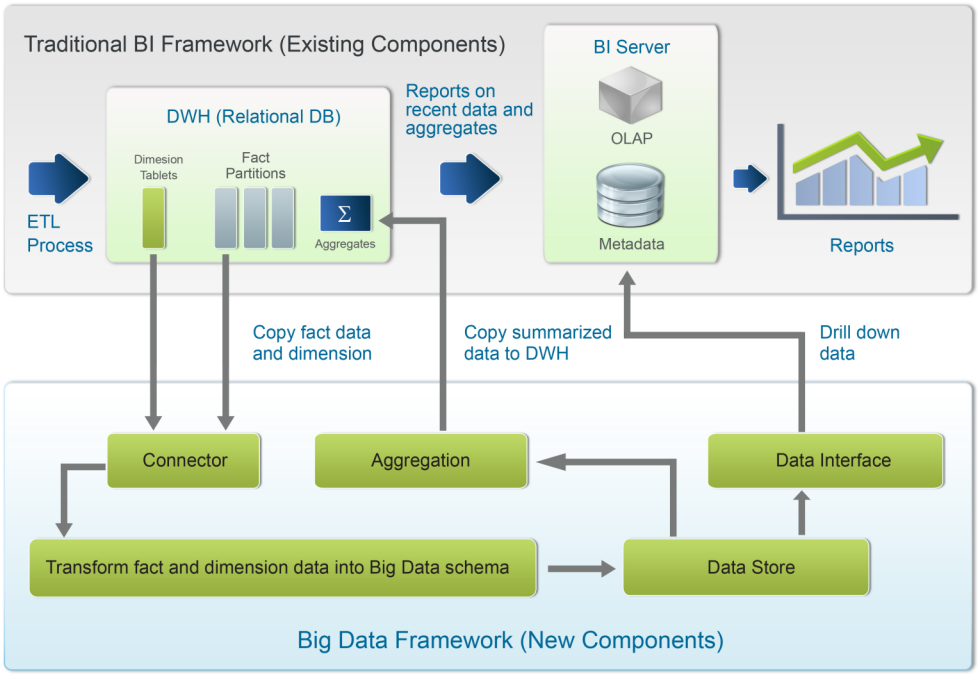
Although many enterprises understand the benefits of Big Data, developing use case based applications on NoSQL can create more data silos, causing deviation from enterprise data integration and metadata management strategies. Big Data lacks strong support for integration with BI systems due to:

- ETL products are still in the process of developing connectors for NoSQL and integrating them with other advanced data transformation operators. NoSQL provides robust and efficient platforms for data processing but the ETL tools lag the capability to push data transformations to NoSQL. This concept is similar to design constructs supported by ELT (Extract-Load-Transform) tools that push transformation and loading operations to relational databases for optimization.
- Noncompliance with SQL standards is a challenge when using Big Data with BI tools. Although simple reports can be generated using Big Data platforms such as Hive that support partial SQL syntax, complex BI features cannot be implemented on Big Data platforms.

Enhancing BI Architecture to Harness the Power of Big Data

A long term enterprise architecture strategy must have strong features from both worlds- DataStack 2.0 and Big Data. As enterprises extensively use relational databases in Business Intelligence and data warehousing environments, the most logical step would be to integrate Big Data with existing data systems to enhance their capabilities. This can be done by using Big Data as a more powerful and data crunching storage system, which acts as a logical extension to the existing data warehouse. The figure below describes the enhanced BI architecture.

Figure 2: Enhanced BI Architecture - BI Big Data Framework



As enterprises extensively use relational databases in Business Intelligence and data warehousing environments, the most logical step would be to integrate Big Data with existing data systems to enhance their capabilities.

An existing Business Intelligence infrastructure can be augmented with Big Data components to build an enhanced architecture by:

Step I - Adding a Big Data Infrastructure to Support DWH

Big Data framework tightly integrates with DWH to pull incremental data from fact and dimension tables. A Sqoop based connector can be effectively utilized to identify changed data in the DWH based on timestamps and copy it to the Big Data tables. The data extraction can be scheduled as a post load activity of the ETL process to sync up Big Data with DWH. With Big Data acting as an online standby system, DWH does not need to store complete historical data, thereby reducing the amount of data stored in the relational database, resulting in optimized data management.

Data partitioning between DWH and Big Data ensures that recently queried data becomes more frequently available on the relational system for various quick ad-hoc analysis supported by BI tools, and detailed historical data can be queried from the Big Data framework for drill-down reports.

Data structures and data types in Big Data are capable of storing DWH fact data, dimension data and relationships. For performance optimization and simplicity of querying on Big Data platforms, de-normalization of DWH schema can be considered.

3 Steps to build reference architecture without impacting existing BI setup -

Step I - Adding a Big Data Infrastructure to Support DWH

Step II - Changing BI metadata to work with Big data

Step III - Implementing Summarization Logic on Big Data

Step II - Changing BI metadata to work with Big Data

Drill-down BI reports and BI metadata will undergo some restructuring to access detailed data stored on Big Data platforms. Minor modifications in the BI Metadata will allow reports to switch between DWH and Big Data in order to optimize report execution by picking summary data from the DWH and detailed data from the Big Data platform. Manual efforts need to be invested to leverage optimal benefits of the enhanced architecture. Big Data framework also provides a robust platform for processing offline reports. Big Data technologies such as Hive and Impala support simple SQL constructs to query data using JDBC connector that can be integrated well with any BI tool.

Although, many NoSQL systems support high performance reads and writes, Hadoop MapReduce paradigm is widely considered as an offline processing framework for very large datasets. However, recent technology developments such as Hive on Hbase and Impala can bypass MapReduce invocations to efficiently fetch data from Hadoop using SQL interfaces. These components are essential to ensure a quick response time for querying detailed data on Big Data.

Step III - Implementing Summarization Logic on Big Data

Post processing steps such as summarization can be pushed to Big Data platforms to utilize its data processing capabilities. This design will be helpful especially for the ELT tools that utilize a database for data transformations, which can potentially lead to degradation in the reporting performance during DWH load process. Any mainstream analytics jobs can be moved to Big Data to run periodic trending, forecasting, and mining; the updated analytics models can then be loaded to DWH for reporting purpose. Big Data technologies such as Mahout provide a rich open platform for text classification, clustering, pattern mining, regressions, and many more algorithms, and can be effectively utilized for this purpose.

The ETL process is a complex data processing module, and remains unchanged with the addition of Big Data framework. The DWH and summary table schema need not be modified, as they are governed by business requirements. Technology enhancements should not have any impact on them.

Benefits of Enhanced BI Architecture

Enterprises have a lot to gain with the enhanced BI architecture that is created by augmenting BI architecture with Big Data components.

- Enhanced architecture effectively reduces cost of ownership by leveraging cutting-edge relational database features and NoSQL. Storing large volumes of data on relational database will impact its performance and also increase maintenance costs. Moving infrequently used historical data to NoSQL will ensure that data is available online for querying, reducing the burden on relational databases.
- Detailed data becomes available in NoSQL for data management professionals to run analytics for discovering patterns, without any impact on normal processing of BI reports. Developers can use any programming language supported by NoSQL to perform niche custom analytics. Insights found in the form of analytics models can be pushed back to a relational data warehouse to create standard BI reports on top of it.
- A Big Data framework ensures high availability of the data warehouse and enterprises can further cut costs in data archival and back-up solutions. Complete historical data can be retained for analysis and audits, without the data needing to expire.
- Existing ETL process for loading data warehouse can be augmented with the processing power of Hadoop based framework to perform post-load activities. These activities include summarization, materialized view refresh, etc.

Integration of relational and Big Data systems will align with enterprise data architecture, rather than creating a data silo for a specific Big Data use case. For certain enterprises that are still pondering about use cases to introduce Big Data and understand its tradeoffs, this architecture will serve as the perfect Big Data kick-off strategy.

Conclusion

Enterprises can enhance their existing BI architecture by leveraging the powerful Big Data platforms on traditional BI systems. The reference architecture can be built incrementally without impacting the existing BI setup by first adding a Big Data infrastructure and establishing a data sync process with DWH. After developers and data analysts are comfortable with data and querying interface on Big Data, BI metadata can be modified for Big Data by creating sample reports to access detailed (drill-down) data. Subsequently, summarization logic can be implemented on Big Data, and the analytics team can start working on identifying interesting business patterns using Big Data mining tools. These jobs can then be moved to a mainstream data processing pipeline to complete the architecture.

We believe BI and Big Data integration brings out the finest analytics capabilities in the industry. More system architects, program managers, and CIOs should exploit this combination to gain maximum ROI benefits from their Big Data rollout strategy.

PERSISTENT

About Persistent Systems

Established in 1990, Persistent Systems (BSE & NSE: PERSISTENT) is a global company specializing in software product and technology services. For more than two decades, Persistent has been an innovation partner for the world's largest technology brands, leading enterprises and pioneering start-ups. With a global team of more than 6,000 employees, Persistent has 300 customers spread across North America, Europe, and Asia. Today, Persistent focuses on developing best-in-class solutions in four key next-generation technology areas: Cloud Computing, Mobility, Analytics and Collaboration, for telecommunications, life sciences, consumer packaged goods, banking & financial services and healthcare verticals. For more information, please visit: www.persistentsys.com.

India
Persistent Systems Limited
Bhageerath, 402,
Senapati Bapat Road
Pune 411016.
Tel: +91 (20) 2570 2000
Fax: +91 (20) 2567 8901

USA
Persistent Systems, Inc.
2055 Laurelwood Road, Suite 210
Santa Clara, CA 95054
Tel: +1 (408) 216 7010
Fax: +1 (408) 451 9177
Email: info@persistentsys.com