INFOCEPTS

Real-time Big Data Integration into Your Existing BI/DW Environment

Technical Expertise

Big Data
Hadoop, Hive, Map-Reduce, Cloudera, Riak, Apache Flume, Cassandra

Data Integration
Informatica, SQL Scripts, Microsoft SSIS, Oracle Data Integrator, Talend

Databases
Oracle, Microsoft SQL Server, IBM Netezza, Teradata, DB2, Greenplum, AsterData, Vertica, Par Accel, My SQL, PostgreSQL, Essbase

Business Intelligence
MicroStrategy, Cognos, Microsoft BI, Pentaho, TIBCO Spotfire, Tableau, OBIEE

Organizations today are confronted with the challenge and the opportunity of data growing at unprecedented rates. This data comes from numerous sources – ERP systems, Data Warehouses, Website logs, Web Services, Social Media, Mobile devices, Sensors, etc. - in various forms - Structured, Semi-structured and Unstructured. “Big Data” is the catch all phrase for this rapidly changing field. Big Data analytics has the potential to provide great insights and opportunities to organizations in the areas of consumer behavior, marketing, fraud detection and customer service. With the right technical architecture, true real-time decisioning is enabled providing organizations with heightened agility. While most organizations recognize the importance and benefits of Big Data analytics, there are challenges arising from the nature of Big Data and limitations of existing technologies that need to be considered.

Big Data Challenges

Volume: With Big Data, the volume of data in storage is a critical issue. Big Data solutions need to have the capability to process Terabytes and Petabytes of raw data. Traditional storage technologies lack the ability to handle the massive data growth originating from Big Data sources.

Variety: Big Data sources like social media, web logs, clickstreams, RFID sensors, etc. generate semi-structured (weblogs, XMLs and flat files) and unstructured data (social media and sensor data). However traditional Data Warehouse RDBMS platforms are best suited for structured data. Designing data structures and establishing relationships with unstructured data adds a lot of complexity to the data models.

Velocity: The frequency of data generation and/or delivery from certain Big Data sources like clickstream data and sensor data is very high and requires users to make business decisions in real time. Traditional BI tools and statistical analysis tools are focused on historical analysis and lack real-time analysis capability.

Variability: As the sources of Big Data are many and varied, it is important that the data used for analysis is reliable, which requires organizations to have a strong data quality validation process and framework. Many organizations lack this. Furthermore, ensuring data quality of real-time data is big challenge.

Integration: The challenge of Big Data integration lies in integrating unstructured data with traditional structured Data Warehouses.

Key Solution Artifacts:

- Solution Architecture document
- Big Data Solution implementation plan
- Data collection System design document
- Big Data quality test case document
- Solution deployment document

- Collection and analysis of high velocity data requires an efficient data integration design that runs rapidly. Traditional, time consuming complex ETL processes do not adequately meet this need
- Extraction and transformation of semi–structured and unstructured data into meaningful structured data is a complex process

Data Visualization: The higher volume of data requires more sophisticated and complex visualizations than the visualizations typically used for aggregated and filtered warehouse data

The existing BI and DW systems of most organizations are limited in their ability to handle Big Data. However, organizations have invested heavily in their existing systems and replacing the entire environment to enable Big Data analytics is typically not an economically viable option. Our solution provides a cost-effective approach to enable existing BI and DW systems for Big Data analytics.

The InfoCepts Solution

Our solution addresses the challenges of implementing a Big Data solution within an existing BI and DW architecture. We use an Agile BI methodology, Data Virtualization and Slow Integration process to implement the solution. Our solution involves adding the following components to an existing Data Warehouse architecture:

Data Collection System: In a traditional Data Warehouse architecture, data is delivered to the ETL staging area either directly or using a file system. In some architectures, an Operational Data Store (ODS) is also used to store structured data. Our solution incorporates a Hadoop/Hive data collection system that can store as well as process data delivered in real time, in batches and as files. These platforms ensure high data quality and faster data integration as they are designed to handle structured as well as unstructured data.

Data Virtualization: To enable real-time analysis, we've incorporated Informatica Data Services (IDS) for data virtualization. IDS virtually integrates all types of data without actually moving the physical data thereby saving the overhead of data integration for all the raw data.

MicroStrategy Visual Insight: Traditional BI platforms are challenging to use for real-time data analysis due to the development time needed to create the required metadata. To enable real- time, ad-hoc reporting, we've incorporated MicroStrategy Visual Insight into our solution. This tool can query the data sources via IDS, allowing the Business Users or Analysts to discover insights from their data and take immediate action (Data Discovery).

Slow Integration: With IDS, raw data is available for analysis as soon as it is delivered to the collection system. With Visual Insight users analyse the raw "big" data to discover data that is relevant, meaningful and useful to them. This is achieved without depending on IT to create business views and merge or cleanse the data. Relevant data once captured, is cleansed and incrementally integrated into the Data Warehouse. This process is called Slow Integration. BI reports are modelled and developed to provide on-going, historical analyses.

INFOCEPTS

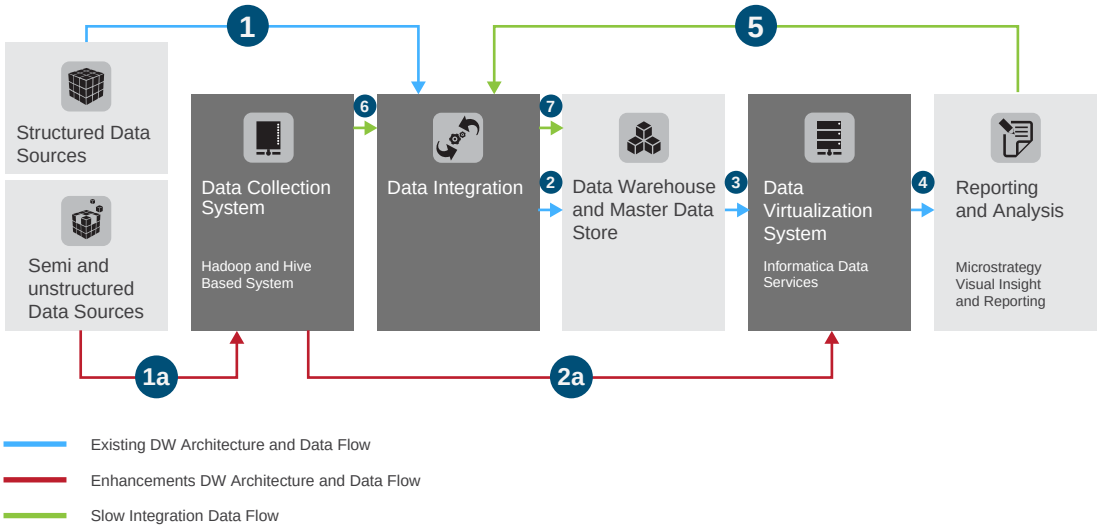
Our Credentials

Our company is exclusively devoted to designing, building and supporting data integration and business intelligence solutions. We combine a process-oriented approach and programmatic best practices deployment with a global delivery model to produce high ROI and quality for our customers. Our BI services span from high quality Mobile Apps, Dashboards to complete end-to-end business intelligence development and support using a host of technologies. Our data integration services include design and development, Big Data integration, Data Virtualization, Data Governance, and Post Implementation Support and Optimization. Customers that we have had the privilege to work for range from larger organizations such as PetSmart, NBC, Nielsen, Toys R Us and Publicis, to a host of small to medium sized organizations.

For more information please contact:

(703) 289 - 5117
sales@infocepts.com
www.infocepts.com

Solution Architecture



Data Flow after implementation of our Big Data Solution

The Hadoop and Hive Data collection system assembles both semi-structured and unstructured data (1a). This system works in parallel with scheduled, routine ETL processes. In our solution, unstructured data is the primary target for real-time analysis. As such, it is fed directly into the Data Virtualization system (2a) where it is used by analysts to discover relationships, trends etc within the real time data feed. Data that is meaningful and useful is then integrated back into the Data Warehouse (5,6,7). With this architecture, our solution preserves the data flow of the existing Data Warehouse system (1,2,3,4) while integrating Big Data analytics into the system.

Why the InfoCepts Big Data Solution?

- The solution has a modular framework – every implementation phase is independent of others
- Our solution delivers high ROI. It seamlessly integrates with your existing DW/BI systems, minimizing the investment required for a Big Data solution
- The solution provides real-time Big Data analytics capability
- It eliminates the need for any specialized ETL skills or user training for Big Data tools
- Flexibility in technology options - the solution is based on a framework that can be implemented using various Big Data, Data Virtualization, Data Warehousing and Business Intelligence technologies