Mobilizing Data for BI and Analytics
Table of Contents

- Mobilizing Data for BI and Analytics .................................................. 2
- Information Architecture: A Capability-Driven Approach .................. 3
- Performance, Latency, Security, And Governance ......................... 5
- Components of Tomorrow’s Data Architecture .............................. 6
- Case Study: Healthcare Payer Leverages Data to Provide Improved Cost of Care at Reduced Risk .......................... 8
- Transforming Information Architecture ............................................. 9
Today, enterprises have access to humongous data for millions of their consumers. This includes a variety of data types. For a healthcare insurance provider, this could be - structured claims data, unstructured HIPAA documents, image files that contain reports and doctors’ prescriptions and voice data as the patient interacts with the customer care, social site data and so on. Most of this data is not used in the decision making process today. Even for the data that does find its way into decision making, the users of the data are challenged by the latency with which their changing requests are addressed by the existing infrastructure.

A data lake-based Hadoop architecture will only partially solve the problem. While it supports the complexity of transformations and provides the capability to support various kinds of data types, it does not support the concurrent ad-hoc requirements of business users.

Traditional architectures have held organizations in good stead, providing the information they need. But in this digital information age, an agile and modern data architecture is required.

Would adding a self-service BI address the overall issue of providing insight to the users? Perhaps. But what about the needs of the business to be able to identify anomalies in claims? This would require a sampled data set for data scientists to create data models and perform deep exploratory analytics.

All of these are distinct workloads, with different capabilities and service level expectations. No single system can meet these differing workloads efficiently well. Modern information architecture is a big step away from the monolithic environment required to transforming to a multi-platform ecosystem that supports timely decision making with an unflinching based on speed driven insights. Adding a lot more complexity to the environment.

A capability-based transformational approach provides the right advantage for creating an agile and flexible ecosystem that can meet the very disparate needs of the organization today.
Information Architecture: A Capability-Driven Approach

The supply chain of data starts at the point at which the data is generated and ends at the point that the insight is delivered to the consumer. As data travels, it goes through a series of architectural components – data ingestion, storage, processing, and finally analytics and discovery. The decision around technology choices for each of these components is driven by the functional capabilities, as well as the ability to deliver the non-functional requirements such as performance, latency and security.

Answers to these questions will determine data integration and governance strategy

What are the different data types – structured, unstructured, semi-structured? What is the frequency with which the data is refreshed? With the advent of public data, data source also becomes an important aspect to understand data quality considerations. How is the new data source mapped to the entities identified as part of the enterprise data store?

Data ingestion architecture would be driven by the sense-and-respond characteristics of the workload

Are these highly critical sensor-based real-time events or near real-time transaction events? Do they require real-time analytics or is there no need for any immediate analytics? Are any of the data entities reference or master data? Responses to these questions will determine if there is a need for batch data processing, real-time event stream processing or data virtualization.

Considerations for the Data Processing Architecture

Do we need to apply minor transformations at the point of data collection or perform complex ETL on an analytical platform? Is the data unstructured and does it require complex pre-processing to a more structured format? Data quality considerations and how those need to be addressed are an important part of laying out the data foundation.

Different groups of users need very different approaches to get to the insight

Data Scientists use analytics sandboxes to tackle complex analytical workloads that could run for days in search for answers to questions. CXOs use existing simulation models to ask “What-if” questions that require the response to be of permissible latency. Business users would like to view the current state of operations – trying to drill down to the data. In which cases, the querying layer needs to ensure sub-second responses. Business processes would require real time recommendations – that work on huge volumes of historical data and high velocity of real time interactions to provide sub-second responses back to the point of interaction.
Accelerated insight from a supply chain is made possible by identifying the nature of the workload and its architectural and functional requirements to determine the components that make up the pipeline.
Performance, Latency, Security and Governance

Performance considerations are driven by the service expectation of the workload. This will determine the usage of MPP components, in-memory technologies, caches, nature of the real-time support. For instance, in a financial industry that is constantly flooded with data, it needs to be captured, stored, and analyzed for patterns and outcomes for petabyte-sized datasets. In-memory computing on GPU based systems can prove to be way more efficient for economically viable performance.

As new data types are introduced in the mix, additional capabilities need to be integrated to ensure that organizations stay compliant to their policies. Public data types or third party data types can help answer some questions better, throwing light on additional dimensions of the entities. However, these data sets are not owned by the organization and would require additional facets – like data anonymization (SSN)/data privacy (social data). In some cases as organizational data travels through the network or is accessed by different users, data encryption and data authorization needs to be enforced. As more data types are added, data security considerations become involved and complex too. The better integrated the security is to the infrastructure, the easier it gets to manage and maintain.

Not everything that can be counted counts, and not everything that counts can be counted.

Albert Einstein, Physicist
Components of Tomorrow's Data Architecture

As organizations adopt new systems such as data lake and self-service BI, it is important that these systems integrate with the existing systems within the environment to ensure synergies.

Below are some of the key components that form a part of the modern data architecture:

**Data Lake**
Stores all data in the format in which it is received and provides economies of scale for the big data problem.

**Architectural considerations:**
- Supports distributed storage using cluster of Hadoop-based nodes
- Supports batch processing of data at scale using MPP at the data node
- Helps developers create specific tools to make data available for users

**Operational Data Stores**
An operational data store integrates data from multiple sources to enable operational analytics. Existing ODS systems don’t scale easily to the volume of data that an enterprise is exposed to today or to the complex analytic needs and the demand for reduced latencies. Therefore, data needs to be periodically archived to retain efficiency. Early adopters are already seeing benefits of analyzing multi-year data to identify early risks and opportunities.

Organizations need a flexible and agile environment to meet the fast pace of the business.

**The environment must:**
- Enable distributed processing and storage – built on Hadoop/Spark clusters
- Improve performance using in-memory technologies
- Support capability to take in data with reduced latencies using Spark-like pipelines
- Support fastest time-to-insight with online data query using query tools like Impala and Drill
- Enforce compliance and secure private data with integrated security tools to control access

**Master Data Management**
MDM systems manage and govern and reference data for the enterprise to create a single version of truth record. These systems merge and match information using data that comes in from various sources to create a synthesized view of an entity. With the advent of new data types, MDM also uses additional behavioral and interaction data collected from social sites and saved in data lakes. MDM is critical to ensure that the data deluge from big data makes sense in the enterprise context. It provides the referential integrity that is lacking in the big data world. Traditionally, it has been a centralized repository of master data across the organization. MDM:
- Creates a comprehensive 360-degree view of the customer based on data collected from external sources. MDM integrates with big data to enrich enterprise-specific data with external data
- Provides order and context to the big data when it is integrated with MDM
- Identifies relationships and influencers across customer segments using Graph database

**Hybrid Enterprise Data Warehouse**
Traditionally, EDW has been the repository of enterprise specific data with capabilities to support reporting, BI, and analytics. Existing EDWs typically don’t scale to big data. As new data sources are identified, the rigid data model makes the time-to-insight really high. Additionally, complex transformations take a very long time to run, delaying responses to requests for any reports or insights.
A monolithic approach to EDWs is significantly cost-prohibitive. Hybrid engineering of information integration solutions is required to leverage the new-age solutions on cloud or commodity hardware in data centers in-house. These hybrid EDWs are a logical approach to the ‘centralized repository’ and support:

- Reduced latency of operations for complex transformations, and accommodate new age attributes that find no place in rigid EDWs
- Real-time analytics on streamed data
- Ad-hoc and advanced analytics – on premise, on the cloud, or in-appliance using in-memory data analytics
- Self-service BI with tools that provide easy access to enterprise data

Integrated with data lakes it creates a more comprehensive view of the data and higher performance on complex transforms when executed in an MPP environment. This allows for reduced latency of operation for complex transforms and accommodation of new age attributes that have no place in the rigid EDWs.

Data Integration
Today’s integration capabilities must support data movement and transformation while dealing with high volumes, high velocity and a wide variety of data. Additionally, the store house of information is not a single physical data warehouse, but a more fluid and logical structure. Data integration needs to be a lot more holistic than the ETL of yesterday. The integration infrastructure must:

- Support the requirement for data integrity - quality, lineage, and governance - considering the information collected would be used to make future decisions
- Use MPP techniques to extract, load and transform data on Hadoop platforms
- Support streaming capabilities of high velocity events without fail and enable real-time analytics with the possible use of in-memory analytic techniques
- Support data virtualization techniques to make reference data available in the information highway
- Enable the application data to participate in the data supply chain by integrating with other applications using standard technologies of extraction and integration
- Support foundations of metadata management, data cleansing and standardization

Business Intelligence
As businesses adapt to change, questions that organizations ask from the data also change. The system needs to be a lot more agile to take in new data sets and provide quick insights using a scalable infrastructure. The infrastructure must provide the responses that users need as quickly as possible using the complete data.

The BI requirements span a spectrum of traditional BI from reporting and dashboards to information discovery, data mining, and data scientist functions in terms of machine learning and advanced analytics. The BI infrastructure must support:

- Refurbished BI tools for collecting data from the new fluid architecture to creating reports and dashboards
- Self-service BI tools to support ad-hoc queries from business users. Business users can then create and share reports and visualizations across the organization
- Sandbox environments for data scientists to identify anomalies and for ideation
- API-based interface that uses applications for business process automation
- BI and analytics collaboration that enables user interaction with reports by including the ability to add comments and annotations, and share them with colleagues
- A Query layer that will allow for high speed complex queries to be run on the underlying data using familiar SQL like tools

Stored annotations can be used to compare data across periods. For instance, if CMOs notice a decline in sales, they can comment on the chart, authorizing a major marketing campaign in a specific region. CMOs can also search for such annotations and use them to compare sales before and after the event.

As data consumers might use a variety of devices for accessing information, the infrastructure must enable optimal visualization for all screen sizes and the ability to support multiple channels.
Case Study: Healthcare Payer Leverages Data to Provide Improved Cost of Care at Reduced Risk

The healthcare payer operates in a complex ecosystem. It is needed to balance its ability to provide the best care possible with rising treatment costs. In addition, recent healthcare reforms have resulted in a more competitive and complex marketplace with a fundamental shift in the way the payer’s products are bought. The company wanted to be able to predict the actual cost of care — one of the most challenging problems in the industry, as overcharging means losing customers and undercharging can result in operational losses and dissatisfied shareholders.

Architecture, technologies, and infrastructure were primarily monolithic and grew over time. They couldn’t handle the scale or the complexity the new business model required. The organization staged the data – which included non-enterprise data like social data – on a big data system. This was integrated with the enterprise data warehouse. The EDW was used to store well-defined entities only. An ideation sandbox was created to help data scientists develop complex models and perform deep analytics. This was used to create specific front end tools for the business user to manage population healthcare spend and predict the high risk members.

A combination of tools and technologies helped provide improved cost of care at reduced risks. This was built iteratively starting with standardization of data warehouse on one side and creating the sandbox to ideate and build models. Subsequently when specific insights were identified, tools were created for business users to allow them to optimize the cost of care for specific patient groups. This incremental, evolutionary and capability driven approach, allowed the organization to see quick benefits at each time and provide the necessary agility that compressed their time to insight significantly.
Large organizations have tried to justify enormous investments in data warehousing systems by trying to put workloads on those systems that are not ideal fits. They are now held hostage to the high costs of operating and slow processes of changing those systems to new problems. Emergence of new compute technologies and strategies like GPU / MPP is likely to dramatically change the data landscape.

While the transformation is about making technology decisions aimed at maximizing value, it is important for the effort to be driven by business priorities to retain a BI focus and to ensure that stakeholders continue to stay invested. Optimal benefits can be realized only with a targeted approach and a business driven strategy, where the initiative starts with a specific business case instead of merely consolidating data and trying to identify nuggets of information from it.

New business drivers or questions will translate to specific workloads that the architecture must now support. This evolutionary approach to building the target architecture mandates an underlying ecosystem that is responsive, flexible, and adaptable. While organizations embark on this transformational journey, a business driven strategy is required, that will translate the business need to specific workloads, and identify the best tools to provide the capabilities to save money, improve performance and increase the speed with which they solve problems.
ABOUT UST GLOBAL

UST Global is a digital technology services company that provides next generation digital solutions for Global 1000 companies. Our mission is to ‘Transform Lives’ using the power of digital technologies and the focus is on digital services and solutions. With a business model of ‘fewer CLIENTS, more ATTENTION’, UST Global strives for excellence in providing our clients with the best service and commitment to long-term client success.

Headquartered in Aliso Viejo, California, UST Global has over 17,000 associates operating in 21 countries across four continents. For more information please visit www.ust-global.com

About the Author

Mandira Shah is a Solutions Architect with the Big Data and Advanced Analytics Services in UST Global and provides big data solutions to major enterprises in their journey of transformation. Her 20 years of experience includes consulting roles across multiple domains for distributed business applications, where the focus is on delivering business capability.

Acknowledgement

The author is grateful to Niranjan Ramsunder – Head, Business Solutions, UST Global Tony Velleca – Chief Information Security Officer, UST Global, Niranjan Ram – Practice Director Business Solutions UST Global for sharing their inputs.