Introduction

Business Intelligence (BI) and analytics provides an essential set of technologies and processes that organizations have relied upon over many years to guide strategic business decisions.

With the emergence of new data sources such as social media, mobile applications, and sensor-equipped “Internet of Things” networks, organizations can extend BI to deliver real-time insight and discovery into such areas as operational performance, customer satisfaction, and competitor behavior.

However, these new data assets, often referred to as “big data,” challenge many of the previous assumptions around data storage and processing for BI and analytics applications. Not only do organizations have to manage much higher volumes of information, the data itself arrives at much faster rates and is more complex and dynamic than existing transactional sources. To be useful, the data must be analyzed and visualized in shorter intervals than traditional reporting cycles.

While many BI systems will be able to adapt to these new requirements, the underlying databases may not afford the same flexibility. Organizations need to explore alternative technologies that augment their systems to fully integrate and benefit from big data.

With its rich document model, powerful analytical capabilities over high volumes of multi-structured data sets, and broadest integration with leading BI and analytics tools, MongoDB provides a foundation to evolve BI to support real-time analytics for big data applications.

The Big Data Challenge for Business Intelligence & Analytics

In traditional BI platforms, the flow of data – starting with its acquisition from source systems through to transformation, consolidation, analysis, and reporting – follows a well-defined sequential process

Operational data from multiple source systems is integrated into a centralized Enterprise Data Warehouse (EDW) and local data-marts via Extract Transform Load (ETL) processes. Reports and visualizations of the data are then generated by BI tools.

This workflow is optimized for users, enabling the deep historical analysis used to inform strategic decision making at senior levels within the organization. Databases and reporting queries are predicated on a number of assumptions:
1. Predictable Frequency. Data is extracted from source systems at regular intervals – typically measured in days, months and quarters.

2. Static Sources. Data is sourced from controlled, internal systems supporting established and well-defined back-office processes.

3. Fixed Models. Data structures are known and modeled in advance of analysis. This enables the development of a single schema to accommodate data from all of the source systems, but adds significant time to the upfront design.

4. Defined Queries. Questions to be asked of the data (i.e., the reporting queries) are pre-defined. If not all of the query requirements are known upfront, or requirements change, then the schema has to be modified to accommodate changes.

5. Slow-changing requirements. Rigorous change control is enforced before the introduction of new data sources or reporting requirements.

6. Limited users. The consumers of BI reports are typically business managers and senior executives.

**Evolving BI and Analytics for Big Data**

Businesses want to harness new data sources and fast time-to-insight in new and compelling ways.

Examples include:

- Retailers tracking user preferences, web clicks, and social sentiments to identify and automatically target geo-aware and device-aware personalized content and promotions

- Utilities capturing household energy usage levels to predict outages and to incent more efficient energy consumption

- Governments detecting and tracking the emergence of disease outbreaks via social media signals

- Oil and gas companies taking the sensor output from their drilling equipment to make more efficient and safer exploration decisions

The availability of new data sources generating big data is challenging the previous assumptions of data management and reporting within the BI platform.
The Need for Speed & Scale

Time to value is everything. For example, having access to real-time customer sentiment or logistics tracking is of little benefit unless the data can be analyzed and reported in real-time. As a consequence, the frequency of data acquisition, integration and analysis must increase from days to minutes or less, placing significant operational overhead on BI systems. In a growing number of cases, source data needs to be analyzed in place in order to provide the responsiveness demanded by the business.

The availability of new data sources drives an explosion in the amount of data organizations must manage, with analysts estimating a doubling in volumes every 12 to 14 months. Not only do BI databases have to handle much higher ingestion rates (often referred to as “data velocity”), there is also the challenge of how data is moved through the BI pipeline, from source systems to the EDW, data-marts and into analytical and reporting processes.

Agile Analytics and Reporting

With such a diversity of new data sources, business analysts cannot know all of the questions they need to ask in advance. Therefore an essential requirement is that the data can be stored before knowing how it will be processed and queried.

The Changing Face of Data

Data generated by such workloads as social, mobile, sensor, and logging is much more complex and variably structured than traditional transaction data from back-office systems such as ERP, CRM, PoS (Point of Sale), and Accounts Receivable.

The existing relational databases used by these back-office systems are designed to model cleansed and neatly structured data into tabular row and column formats with defined values, enforced by rigid schemas. They were never designed for the polymorphic, semi-structured or unstructured data that is now typical in many of today’s big data applications.

Higher Uptime Requirements

The immediacy of real-time analytics accessed from multiple fixed and mobile devices places additional demands on the continuous availability of BI systems.

Batch-based systems can often tolerate a certain level of downtime – for example, for scheduled maintenance. Online systems, on the other hand, need to maintain operations during both failures and planned upgrades.
Taking BI to the Cloud

The drive to embrace cloud computing to reduce costs and improve agility means BI components that have traditionally relied on databases deployed on monolithic, scale-up systems have to be re-designed for the elastic scale-out, service-oriented architectures of the cloud.

Impacts to Traditional BI Databases

The relational databases underpinning many of today’s traditional BI platforms are not well suited to the requirements of big data:

• Semi-structured and unstructured data typical in mobile, social, and sensor-driven applications cannot be efficiently represented as rows and columns in a relational database table.

• Rapid evolution of database schema to support new data sources and rapidly changing data structures is not possible in relational databases, which rely on costly ALTER TABLE operations to add or modify table attributes.

• Performance overhead of JOINs and transaction semantics prevents relational databases from keeping pace with the ingestion of high-velocity data sources.

• Quickly growing data volumes require scaling databases out across commodity hardware, rather than the scale-up approach typical of most relational databases.

Relational databases’ inability to handle the speed, size, and diversity of rapidly changing data generated by modern applications is already driving the enterprise adoption of NoSQL and Big Data technologies in both operational and analytical roles.