Industry Context and Trends

In the digital age of social media and connected devices, enterprises have a plethora of data that they can mine, to discover hidden correlations and perform deeper real-time analytics in order to deliver better customer experience, improve operational efficiency and monetize their data assets to uncover new revenue streams.

In addition to answering the traditional questions around sales, performance, profitability etc., enterprises are now looking to discover correlations between seemingly unrelated streams of data to gain a competitive advantage.
Discovery of such hidden correlations and insights require platforms that
- Ingest, Store and Process large volumes of high velocity in real time from a variety of data sources
- Generate rapid insights by co-relating multiple structured and un-structured data streams
- Enable data scientists to discover correlations and model algorithms that can deliver insights
- Self-learn based on new data patterns and improve accuracy of insights

With the advent of new types of data emerging from web and connected devices, the challenges in delivering these insights have increased multi-fold and put a tremendous pressure on traditional Business Intelligence platforms.

- **Structured data** grew by more than 40% per year

Types of **Traditional content** including unstructured data is growing by up to 80% per year

An estimated 2.8ZB of data in 2012 is expected to grow to 40ZB by 2020. 85% of this data growth is expected to come from new types; with machine-generated data being projected to increase 15x by 2020. (Source IDC)

This new form of data poses unique challenges for business and IT such as
- Effectively storing large multi-structured data
- Capturing high-speed data and processing it in the right time
- Creating flexible yet high-performance data structures to answer new business questions
- Creating a platform which will provide an integrated and unified access to all the information

**Market Offerings and Gaps**

To address the challenges mentioned above, the industry is now looking at platforms to support the Volume, Velocity, Variety and Value needs of Next-Gen Business Intelligence.

These platforms are typically a combination of Data Platforms and Discovery/Analytical Tools.

Current offerings available in the market can be classified as

**Data Platforms**, which offer open source platforms for building and managing Data Lakes and Big Data Warehouses. These platforms leverage components from open source Hadoop stack and provide platform administration, governance, security and basic discovery capabilities on top of the core stack. Some of the leading vendors in this space include Cloudera, HortonWorks and MapR
**Analytics Platforms**, which offer point analytical solutions to industry use cases such as Customer Analytics, Network Analytics etc. These offerings provide pre-built analytics on top of Hadoop ecosystems to serve specific business needs. Some of the leading vendors in this space include Datameer, Platfora, Guvavus etc.

**Augmented Appliances**, which offer high speed appliances on top of Hadoop storage to speed up data access. Teradata Aster, EMC Greenplum, SAP HANA, HP Vertica and IBM Big Insights are some of the leading vendors in this category.

**Data Discovery tools**, which provide exploratory capabilities on top of Big Data storage platforms. They help ingest data from a variety of data sources, model it and create consumable data sets out of the underlying data. Tableau, QlikSense, Tibco Spotfire etc. are some of the leading vendors in this space.

While these offerings help in addressing some of the challenges posed, they do not offer a silver bullet to solve all of the Big Data Challenges.

<table>
<thead>
<tr>
<th>BENEFITS</th>
<th>Analytics Platforms</th>
<th>Augmented Appliances</th>
<th>Data Discovery Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low cost</td>
<td>Point solutions that come with pre-built analytics</td>
<td>MPP and In-memory capabilities offer fast response time for on-demand analytics</td>
<td>Enable data scientists to access raw data and discover insights</td>
</tr>
<tr>
<td>Simplify the platform management functions and allow organizations to focus on data</td>
<td>Industry standard algorithms</td>
<td>Cuts down on data preparation time significantly</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GAPS</th>
<th>Data Platforms</th>
<th>Analytics Platforms</th>
<th>Augmented Appliances</th>
<th>Data Discovery Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>No pre-built analytics</td>
<td>High cost</td>
<td>High Cost</td>
<td>No data platform</td>
<td></td>
</tr>
<tr>
<td>Minimal discovery capabilities</td>
<td>Point solution that caters to specific use cases and is not meant for integrated analytics</td>
<td>No pre-built analytics</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimal discovery capabilities</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Infosys PoV**

Infosys believes that the challenges posed by Big Data need a “High Performance Analytics” (HPA) platform that provides a comprehensive set of building blocks to provision data, define storage structures, create data sets of consumption, enable exploration and run analytical models against the data.

The solution should offer enough flexibility to extend the available analytical models to suit enterprise specific needs.

The core of any HPA platform is a data management platform that can practically store unlimited amounts of data of any format, schema and type, that is relatively inexpensive and massively scalable.

Data Lakes are designed to offer this capability.

![Figure 2: Logical Architecture of Data Lake](image-url)
They help in two ways

- Information Discovery/Agility in Analytics – simplify the data acquisition to initiate discovery on raw data by exposing business users through discovery tools.
- Data Warehouse Expansion – helps in expanding data warehouse to capture data at a lower grain and higher diversity, which is then fed into upstream systems.

Unlike traditional relational databases, data can be stored in the raw format where analysts and developers can then apply a structure to suit the needs of their applications at the time they access the data with “Schema on Read” instead of “Schema on Write”.

Intelligent Data Discovery tools enable data scientists to build data models and views that can be used for the analysis of structured and unstructured data. They offer capabilities to search on metadata and create data sets for running analytics. They eliminate the need for IT involvement and reduce the time involved in data preparation.

Analytical modules built on statistical tools, enable data scientists to build algorithms and models that can deliver predictive and prescriptive insights. While the analytical models will differ from enterprise to enterprise, a best-in-class analytics platform should have the basic building blocks such as Sentiment Analysis, Text Mining, Fault Prediction, Fraud Detection, Risk Analytics etc. which can be extended based on the enterprise’s specific needs.

Data Lakes combined with Data Discovery tools and Analytical Algorithms form the core of a High Performance Analytics Platform.

High Performance Analytics Platform addresses the challenges posed by Big Data by providing

- Data Lakes built on commodity hardware that are cost-effective for storing large volumes of data
- Distributed processing architecture that efficiently processes large volumes of structured, semi-structured and unstructured data
- Horizontal scalability that can support future needs
- In-memory processing engines to deliver rapid insights
- Pre-built analytical models that can be extended to enterprise specific needs, reducing the time-to-insights
- Data discovery and deep analytics capability that can uncover hidden correlations and deliver deep insights based on ‘all’ available data assets
Success Stories

- Implemented a Business Data Lake for an Australian telecom major to provide insights into various lines of business like Revenue Assurance, Marketing etc. and improved the ability to enable 11.2% of the total USD 26.8 billion revenue persistently, with coverage accelerated to 27%

- Implemented a Data Lake for a leading financial major in US which gave a 360° view of the wholesale customer that improved prospecting effectiveness, market segmentation and positioning

- Increased ARPU, reduced customer churn and identified new revenue streams by selling anonymized data to advertisers and retailers for a Singapore Telco by building a Big Data enabled customer analytics platform

- Created a model to predict ATM failure with 80% confidence level for over 8500 ATMs which resulted in a 14% increase in call center efficiency and 18% cost reduction