ClearStory shares its SaaSy tale of multi-source discovery and analysis


When we last caught up with ClearStory Data, it had recently secured a fresh infusion of capital and was gearing up for a full early-access program in the second quarter, having previously provided its wares in restricted-access mode. The two-year-old startup – founded in late 2011 by key players behind Aster Data, which was acquired by Teradata in March 2011 – has now formally launched, and claims a paying customer base in the double figures, following the completion of early-access trials with enterprises in the consumer packaged goods, food and beverage, financial services, retail, media and entertainment sectors.

The 451 Take
ClearStory's ability to handle a wide variety of data sources and analyze and visualize them in a way that it is easily consumable by nontechnical users holds great appeal and is a strong differentiator when used in tandem with a front end built with easy navigation in mind. The startup also now has a roster of reference customers to provide validation. However, we're still not sure how well it scales, even though ClearStory has a tiered pricing model geared up to 1,000 users. Furthermore, we think the option of a private cloud deployment model, which is technically possible, would increase its appeal by assuaging any lingering security concerns.

Context
ClearStory Data has officially launched its Data Intelligence BI platform and application, which is delivered as a multi-tenant service and hosted on Amazon EC2, although it is not exclusively crafted for it. The SaaS offering is designed to provide a user-friendly discovery and analysis experience primarily to non-technical users who need insight into a variety of different types of information for analytic purposes.

The startup uses the term 'intelligent data harmonization' to describe the smarts that go into integrating a variety of different types of information required for business-oriented analysis. It provides point-and-click access to public and premium data, including economic data from the IMF or US Federal Reserve, as well as sales and marketing information from Dun & Bradstreet and IRI. The data discovery, visualization and analysis service also has native APIs to salesforce.com and Omniture, provides access to flat files, and can hook into publicly available open APIs to offer access to Web-based data sources such as weather information, financial data and social media. It can also integrate relational data and semi-structured and structured data in the Hadoop open source data-processing framework. However, Hadoop and relational data source access requires involvement from data stewards, and is not self-service, like the rest of the sources that it provides access to.

So how exactly is multi-source data integration and harmonization achieved? ClearStory's so-called Data Interference Engine is one lynchpin. As its name suggests, the engine essentially infers semantics from the data being ingested into it by recognizing whether the data is geographic, time-based or currency-based in nature. Once the semantics of the data have been gleaned to ascertain its structure, shape and content, data harmonization occurs in another layer of the BI stack. The data is harmonized on common dimensions that users can choose to keep – or not. The converged insight is then presented to the end user in a so-called story. A story is a visualization crafted from the convergence of multiple sources of
data that is interactive, and is designed to be refreshed from source and continuously harmonized to keep it alive.

The startup also makes use of distributed in-memory processing once the semantics have been garnered from the data to make the subsequent processes of data convergence and harmonization fast. The Data Convergence Engine makes suggestions on how the data should be harmonized based on inferences – e.g., should it be harmonized by time, geography or categorical information? ClearStory makes use of iterative low-latency processing to stream data to the end user depending on the nature and frequency of the data, so if the source data is updating every few milliseconds it is designed to stream at that rate. The BI platform is also designed to reconcile time gaps in various sources of information by performing harmonization on time attributes and adjusting to the different rates at which data is brought into it. Other key components within the underlying BI stack include a data catalog, which might otherwise be called a business glossary as it holds and maintains real-world data definitions; data lineage services; and user management, collaboration and permission services.

ClearStory employs the Storm open source distributed real-time computing environment, which is used for processing streaming data, as its data-ingestion pipeline. That said, it has modified Storm with a custom domain-specific language so that it can handle relational and other types of structured data. The company also uses the Spark open source system for large-scale data analysis on clusters in order to provide low-latency processing to make data exploration fast. In addition, it has modified Spark so that it will interoperate with the BI platform's data catalog and model. The Hadoop Distributed File System is employed as a repository once data has been ingested into the analysis and visualization service, which also has its own language that is designed to make intelligence queries off the back end to create the visualizations.

Over on the front end, there’s a BI application that is designed to be interactive, user friendly and highly visual. Users log in and are provided with a homepage with various tabs, including MyStories and MyData. A story is essentially a piece of visual analysis created by the convergence of one or more data sets, which also includes calculations, filters and other types of operations. Users can author stories, view colleague's stories, invite coworker's to view stories and collaborate on stories, depending on access permission and privileges. They can also bookmark them and send a link of the bookmark in an email.

Stories are created by clicking on the MyData tab, which displays a palate of data sources that the user can pick from to create the visualization. There is also a ‘data you may like’ feature that provides recommendations on information sources that the user may want to bring in for analysis, including public and premium information sources. It is essentially a marketplace that ClearStory has worked with various data providers to create to make data from these providers easily discoverable and usable. Premium data is purchased through a 3- or 12-month subscription, which is currently provided as a line item in the customer's monthly subscription but will ultimately be handled by a billing system, which the startup plans to build into the service. Premium data can be shared with up to five people.

**Competition**

ClearStory Data is one of a growing group of vendors seeking to provide integration and analysis of diverse data types for exploratory and visual analysis purposes. That said, we aren't aware of any other BI player that brings proprietary and public data sources into the analytic mix in quite the same way, although we suspect that there are moves afoot from other firms in the sector to embrace a similar concept.

If we were going to make a comparison between ClearStory and any other company, it would be Tableau Software due to their shared heavy emphasis on user-friendly exploratory and visual analysis – without the need for the user to model the data. However, ClearStory has a stronger multi-source analysis story than Tableau, which is gradually embracing semi-structured information via partners, but was originally founded to visually analyze and discover structured data. ClearStory is also targeting a different target market than Tableau, which continues to employ a land-and-expand strategy to move from midmarket and departmental enterprise deployments into enterprise-wide rollouts. In contrast, ClearStory is targeting enterprises from the outset and focusing on deployments of up to 1,000 users.

ClearStory's support for Hadoop-based analysis also continues to elicit comparisons with Platfora, Datameer and Karmasphere. However, these fellow startups are targeting businesses that have installed the open source data processing and want to get value from analyzing data within it. ClearStory considers Hadoop one data environment requiring analysis, even though it does use Hadoop as part of its BI platform.

Although embracing different types of data than ClearStory, Actuate and Datawatch share a similar goal of bringing information that hasn't previously been analyzed – but holds valuable insight – into the analytic mix. Actuate serves semi-structured data management needs via its
acquired Xenos portfolio, which includes a content and rules engine designed to capture, process and route a variety of structured and semi-structured data such as documents, statements and invoices, as well as data transformation software to map and transform semi-structured and structured data. Datawatch bought Panopticon in June to add an exploratory visual analysis component to its underlying capture and transformation layer for semi-structured and structured data.

Glassbeam is another startup in the structured and semi-structured analysis fray. It shares a focus on a SaaS delivery model with ClearStory. However, the comparisons end there, as Glassbeam focuses on granular machine data analytics. Pentaho is also seeking to unite the previously separate worlds of structured data with new emerging sources of data in Hadoop and NoSQL databases – but is using an ETL and visual-modeling approach, which is a slightly different focus. Jaspersoft is bringing new types of data into the BI mix by providing native reporting on NoSQL databases, which aren’t a focus for ClearStory right now – but could be in the future.

Finally, analytic heavyweights IBM, SAS Institute, Oracle and SAP can also bring some of the same capabilities to bear as ClearStory for the analysis of unstructured and structured data, albeit using multiple products – and consulting services in some situations, as well.

### SWOT Analysis

<table>
<thead>
<tr>
<th><strong>Strengths</strong></th>
<th><strong>Weaknesses</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Providing user-friendly exploratory and visual analysis on data in multiple structures, formats and environments is an analytic nirvana that ClearStory appears able to meet. The startup also now has paying customers to back up its technological claims.</td>
<td>A SaaS delivery model on Amazon Web Services is not for everyone. We'd like to see more proof points of true enterprise scalability.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Opportunities</strong></th>
<th><strong>Threats</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Moving downstream into the midmarket will open up a new revenue avenue with an audience that is a solid adopter of SaaS BI.</td>
<td>There are an increasing number of players looking to serve up analysis of diverse data types, including big guns IBM, SAS, Oracle and SAP, plus commercial open source BI providers Pentaho and Jaspersoft. Tableau is also stretching its reach into semi-structured data.</td>
</tr>
</tbody>
</table>