October 2015

Context

Once upon a time data was neatly arranged into rows and columns. It was a natural format for the highly structured data that accompanied most administrative functions within the business. Sales orders, employee data, purchase orders, accounting information - each transaction was exactly the same format as all other transactions. It made life relatively easy, and database designers could create descriptions of the data (schemas) that were set in concrete, with very little change as the years ticked by.

Nearly all data served the purpose of helping businesses become more efficient, and the focus was firmly on the bottom line. Very few businesses considered that IT might be used for anything other than labour displacement and cost reduction. My own research shows that just a few years ago only one in ten business managers considered that information might be used for top line growth. In fact there was an unwritten protocol that such considerations might be nothing more than flights of fancy. If senior managers in your business still think this way it might be time to jump ship, because businesses in almost every industry are starting to use data in a wholly different way.

Today data is gushing everywhere - social data, data from devices and sensors, text data (customer comments, documents), click-stream data from web sites - and so on. It doesn't mean that all this data is useful. Much of it isn't, but some of it is very useful indeed, and the new science of data is concerned with teasing out the data that might lead to a new market opportunity, result in happier customers, spot a key product enhancement, and a thousand other things which could contribute to top line growth.

This change of emphasis needs a change of mind set. Administrative functions are prescriptive in nature and business intelligence platforms have served to dish out regular reports, and more recently various charts and dashboards. These can tell a business user what has happened or is happening, but they cannot disclose why these things have happened. For that we need causal analysis, and not the descriptive and diagnostic analysis associated with business intelligence (BI).While regular reporting and dashboards with KPIs are as prescriptive as the processes they describe, exploring data for the nuggets that might enhance customer relationships, or reveal a product opportunity, is an iterative, exploratory process.

Now if the only data resource we had was our traditional transaction oriented data there might be little opportunity to significantly enhance the top line. Fortunately businesses increasingly have access to much richer sources of data, and these are data which add wholly new dimensions to understanding customers and markets. Text data in the form of customer comments is a rich source of sentiment, and with the relevant tools can reveal insights that might not be gained any other way. Location data freely transmitted from mobile devices can give businesses an opportunity to communicate with customers when they are near to a store, or a gas station, hotel, airport or any other location sensitive business opportunity. Connections between customers and particularly those revealed in social networks can be analysed for relationships using graph databases. Streaming data from sensors and

devices, both domestic and industrial, allow businesses to anticipate problems and opportunities. Clearly this is something different, and we increasingly find ourselves dealing with data that is much more unstructured, which in turn denies us the opportunity of assuming the data will neatly fit into existing data stores.

The net result of this is that data are becoming more complex, diverse, more urgent and of course being generated in much greater volumes - enter big data.

Big Data

It is increasingly acknowledged that while the term 'big data' is one of the most successful technology buzzwords ever invented, data does not have to be big to qualify. In fact it is diversity of data that is generally much more important when we want to understand customers better, or seize market opportunities in a more timely manner.

The term 'big data' arose naturally from the development of the Hadoop platform - expressly designed to handle very large volumes of data using low cost infrastructure. Companies like Google, Yahoo and Facebook needed to store petabytes of data in a scalable manner, and were the main innovators of big data technology. Many businesses however do not necessarily want to store petabytes of data, but they do want to process greater diversity of data types, and process it in near real-time. In typical fashion, the people who create technology cliches came up with the three Vs of big data - volume, variety and velocity. Some added others (also starting with V) such as veracity. If we were only dealing with volume the problem would almost be trivial - set up Hadoop infrastructure and just keep pumping the data in. The much harder problem is that of variety. We now have protocols for data transmission and specification that support and infinite variety of data structures (XML and JSON for example). The good news is that we can capture complex data. The bad news is that we need to mean sense of it and mould it into a form that is amendable to analytics activities - and it is this requirement that is being met by a new generation of data preparation platforms.

At the present time big data, and especially the Hadoop ecosystem of technologies, is something of a 'do it yourself' assembly kit. This is exacerbated by the continuing rapid evolution of various components, and users of big data technologies need a considerable skill set to both implement a big data installation and keep up with developments. At the current time it is estimated that the total number of organisations using big data in anger number in the low thousands - three or four thousand. Big data will not become mainstream until it becomes more a 'plug-and-play' infrastructure. This will happen, but it is difficult right now to say when and how.

Data Wrangling

So our 'big data' infrastructure will be handling much greater diversity of data. It is no longer simply a matter of connecting to a set of well structured data sources and displaying reports and charts. The data has to be shaped for analysis. Without getting too technical, much modern data may be hierarchical in nature, or formatted on a record by record basis. Analytical tools just cannot make

sense of data in these formats. So the term ‘data wrangling’ has been introduced to cover the processes that are necessary to prepare data for analysis. There are three quite well defined steps needed to take structured, semi-structured, and unstructured data and make it fir for purpose:

Exploration - obviously the first thing to do is determine exactly what is in our data stores. With structured data this is easy - the schema will tell us. With semi-structured data we need to extract the ‘schema’ as the data is being explored. The end result should be a description of the data (metadata) and the nature of the content.

Curation - is the process of building a structure that can be used by analytical processes once th exploration has disclosed the nature of the data.

Production - should the resulting data prove to be useful it may be desirable to production line the creation of the resulting data sets. A host of additional issues include robustness of processes, security, and the automation process itself.

Various skills are involved in this whole process including the data analyst, database administrator, data scientist, and of course the end user. All three steps are driven by need, and data may be needed for BI, data mining and/or data visualisation purposes.

In the era of pure structured data much of this task was accomplished through extra, transform and load (ETL) tools. With diverse semi-structured data types this needs to be augmented with profiling.

Accessing data sources in the first step in the technical process. Many contemporary data preparation platforms come with a large number of connectors to big data sources, traditional relational databases and external applications and cloud based data sources.

The next step is transformation of the data, although transformation and profiling are iterative. The aim of transformation is to create a usable schema and variables that are relevant to the task in hand. This process is complete when data quality is adequate and the data are well described by metadata. The ideal scenario is to have records that are self-contained, in that all the information relevant to an event is contained in a single record. Data can also be enriched with data from other data sources, and of course cleaned for erroneous values, missing values and other discrepancies.

Profiling is largely concerned with creating descriptive statistics. An example might be the distribution of values in an order value field. In fact there are two types of profiling - type based, where the percentage of values in a field that meet the specification for the field is indicated, and distributional where deviation based anomalies are identified.

This is not a trivial undertaking, and while we all want to see dashboards and use algorithms to find useful patterns, the bulk of the work (currently estimated at fifty to eight per cent of the analytics task) is data preparation. And so it should be clear that technologies that might automate this process, partly or completely, should be of considerable value.

Data Preparation Platforms

A number of alternatives present themselves for data preparation. At the most basic level programming languages can be used to transform data, and R, SQL, python and SAS are good examples. Even the Unix shell contains commands that can be used for data transformation, although none of these can be called self-service. Others might choose to use specialised spreadsheet applications, although these do not scale particularly well, and a considerable amount of skill is required.

Traditional ETL tools now often come with a visual interface where boxes that perform various functions are dragged on a canvas, and connections are made between these functions to create a workflow. This is definitely a step up from writing program code, but these tools generally offer little added intelligence to aid productivity.

Fortunately a new generation of data preparation tools is available that do contain intelligent functions to considerably speed the data preparation task. These often employ big data technologies, and specifically Spark with its associated machine learning tools, to automate many data preparation tasks. These are typically visual in nature with rich profiling information and instant feedback on the effects of various transactions. They also tend to suggest transformations based on the nature of the data being processed, and will often join data sources together based on common features.

ClearStory

ClearStory provides a cloud based platform that takes business users from data to visualisation with minimal need for technical skills. In common with several other suppliers (Tamr, Paxata, Platfora and others) it uses machine learning techniques and Apache Spark in-memory processing, to take data from its raw state, to a state where business users can create the data visualisations they need. As data sources become more diverse so businesses are using the Hadoop big data platform as a ‘data lake’, and ClearStory increasingly supports this and many other diverse data sources.

Users are presented with a collaborative environment where data discovery, exploration and visualisation efforts can be shared via StoryBoards. These provide one or more visualisations in the context of a story line, and various users can annotate and comment as the StoryBoard evolves.

Terms such as data wrangling, data blending, data curation and governance are increasingly used to describe the effort that is needed to handle complex data, and while we all want easy to use data visualisation tools, it is the data preparation, and all that goes with it, that determines the success or otherwise of data visualisation. To this end ClearStory is very well positioned for the growing need to handle complex data environments in a data visualisation context.

The users interface provides contemporary drag and drop graphical environment for the creation of large range of visualisations, including maps, charts, graphs, tables and other artifacts. These can be assembled into dashboards, which in turn can be incorporated into a StoryBoard. Context is always maintained, and especially data context, so users can know the profile of data used for a particular

visualisation. As users select the data they want to visualise ClearStory will suggest visualisations, which can be accepted as-is, or modified as needed.

It is well known that data preparation consumes the largest amount of time and effort in any data analysis task. ClearStory exploits machine learning methods in a Spark environment to automate much of this effort. This is not a 'black box' and users are guided as data is prepared for use. ClearStory will infer what data means (identify date fields, dates, numeric data – and so on), and combine data from diverse sources based on its understanding of the data.

Data sources range from Hadoop through to relational database, files and external data sources. These latter are often used to enrich and augment data, and ClearStory is particularly capable of blending open data sources into an analytical task. Such external sources might include census data, business registrations, field survey, media and market intelligence and macroeconomic data such as GDP growth. A 'Data You May Like' library of curated resources means users can direct access to various data sources.

Lavastorm

Lavastorm technologies are truly interesting and innovative. This is not another data visualization or ETL toolset, but a platform that enables a complete workflow from data connection, through data wrangling (merging, cleaning, transforming, profiling), to the creation of visualizations and analytical models, and last but not least the deployment of those models in a production environment with monitoring and alerts.

The Analytic Engine provides a graphical drag-and-drop interface that starts with connections to data sources, and ends with a data visualization or analytical model. These models can be run against real-time data, generating alerts when thresholds are breached, or simply allowing management to monitor performance.

Lavastorm can be used to prepare data for other analytics tools, including Tableau, Qlik and Spotfire. With the data preparation workload accounting for anywhere between 50% and 80% of the analytics task, any technology capable of reducing this overhead, while improving governance and transparency, has to be welcome. Lavastorm claims orders of magnitude reduction in the time to prepare data, and most importantly, the toolset can be used by business users instead of ETL specialists. While Lavastorm has been around for well over a decade, its technologies have suddenly become very timely with the advent of big data and increasing data complexity.

The Analytic Engine is a graphical environment for creating data workflows using a drag-and-drop interface. Lavastorm claims, and it seems fairly apparent, that business users can create their own data flows - and quite complex ones at that. It easily handles the mashing of data from multiple data sources, and takes the user through the processes necessary for cleaning, profiling and transforming the data.

For predictive and statistical analytics Lavastorm offers an enterprise grade R (the open source analytics language) environment. IT can provision sandboxes on the fly, data scientists can compile algorithms into accessible, reusable analytic building blocks, and analysts can deliver insights through the self-service, drag-and-drop predictive analytic functionality to build complex data models and provide validation for their current visualization tools.

Lavastorm Resolution Center - such a dull name for an exciting piece of technology. The Lavastorm Resolution Center (LRC) is where the rubber touches the road. This is where the analytics become actionable, and the LRC comes with integrated alarms, case management, reporting and query capabilities. The uses are numerous and include revenue assurance, retail loss prevention, process improvement and defect resolution programs.

The alarms can notify threshold violations or data anomalies, and controls can be adjusted as business conditions change. In a case management environment alarms can be detected that need follow up. A routing engine allows managers to define rules that automatically create cases, append alarms to new or existing cases, and identify potential suspects. The system automatically adds reference data to suspects in each case, ensuring the right level of problem detection.

Configurable queries, metrics and reports provide analysts with the information they need to focus their research and resolution efforts, and management with the detailed insights they need to understand operational trends and performance. Ad hoc visualizations, pre- defined reports, and multi-report sets that act as basic dashboards are possible.

With Lavastorm, analytics can be shared using existing corporate policies with non- technical consumers via read-only applications that visually display the underlying business logic, the how and what of the results, on any web browser. Consumers can re-run the analytic application using dynamic parameters for ad-hoc requests. Users can also be exposed to all of the analytic applications in the directory and request permissions to gain access paving way for governed data discovery.

Paxata

Paxata provides the tools to significantly speed up data preparation, and offers a contemporary solution that employs a big data infrastructure and automated techniques which exploit machine learning methods. The net result is a self-service data preparation platform that can be used by business analysts and skilled business users to considerably speed up the data preparation task.

The core capability of Paxata leverages Hadoop and specifically Spark, so that large scale in-memory processing is available for the machine learning algorithms that give Paxata much of its power. Paxata can be deployed on premises or accessed as a cloud service. The on-premises deployment requires a Hadoop environment (either dedicated or shared).

When Paxata processes a data source it automatically identifies many data types (dates, products, places etc.). It also provides mechanisms for the rapid identification of data quality problems by engaging in completely ad-hoc interactive exploration with full-text search, interactive text and

numeric filters and histograms, and visual data quality heat maps that highlight patterns, errors, duplicates and sparse or missing data. The central feature of Paxata is it's machine learning based IntelliFusion. This highlights inconsistencies, gaps, duplicate data so that analysts can fill in blanks, remove or rename duplicates, fix inconsistent capitalisation and other tasks needed to improve the data. IntelliFusion's proprietary semantic fusion and machine learning engine, automatically detects common attributes across multiple data sets, then provides best-match options to the analyst who chooses which combination makes the most sense for their analytic needs. Paxata makes AnswerSets available directly through ODBC LiveQuery to Qlik, Tableau, Excel and any other ODBC-compliant analytics tool or application. Paxata also supports publishing AnswerSets to Hadoop clusters.

The formatting of data for analytical purposes is straightforward and data can be pivoted or de-pivoted, columns can be split, and aggregations can be created to quickly make the data sets more suitable for the required analytic exercise. Enrichment, often from third party providers (e.g. zip codes, industry codes etc.), is enabled through the Paxata Library.

Paxata supports a wide range of data sources, including HDFS, relational databases, Excel, Flat Files, XML, JSON and Avro. These data can be integrated into an answer set as needed – this is supported by Paxata's Adaptive Data Preparation capability. In practice it means an iterative approach can be taken to data preparation, as business analysts develop a feeling for the data they need (and don't need).

Data sources are categorized into four types:

- Local Desktop Files – sitting on a local system. These can be uploaded into the Paxata platform as needed.
- Remote Data – is data available over a network, local or remote. By supplying a network path and relevant passwords Paxata connects directly to the data source.
- Databases – these are accessed through a JDBC connection.
- Web services – including apps such as Salesforce.

Once the data sources are known to Paxata, its IntelliFusion functionality will look for relationships between data sources.

Paxata makes use of Apache Spark technology – the in-memory, parallel processing platform that brings speed and flexibility to Hadoop (although Spark can also be used without Hadoop). Paxata is actively engaged in the development of Spark and has developed its Domain Specific Language (DSL) for data preparation tasks.

Platfora

Platfora provides an end-to-end big data data discovery and exploration platform that starts at data ingestion and ends with visualisation. In many ways it is the right product at the right time. Had Platfora tried to deliver its platform just four or five years ago, it would have been shooting at a moving target, since big data technologies were very immature. Today however the growing use and acceptance of Spark in-memory processing, in addition to a maturing of Hadoop, means that Platfora can deliver massively scalable data exploration and discovery tools that overcome many of the problems associated with traditional data warehousing platforms. These are built with predefined needs in mind, using limited data sets, and so are inflexible and slow to design, build and use. Three month latency between need and capability is common with these platforms.

Platfora uses the Hadoop distributed file system (HDFS) as a data store, ingesting data from transaction based systems, devices, external data feeds, and so on. Data is catalogued and prepared using Spark machine learning and in-memory processing. What this means in practice is that users get to see the connections between data sources, and data that has been prepared for analysis.

The Platfora architecture sits on top of the Hadoop platform and provides scalable in-memory processing to handle large queries at speed. Much data preparation is handled automatically as data is taken from its raw state and moved to a structured columnar database within Platfora. Users get to see samples of the prepared data and can override faulty interpretations. Platfora can then create ‘Lenses’ or data marts on an ad-hoc basis as needed – which is of course the silver bullet many business analysts are looking for.

Given a suitable lens, Platfora can then be used to visualise data as required. To this end the Platfora Vizboard engine supports visualisation of millions of data points, with the zooming, drill down, panning and filtering functions typical of a product of this nature.

This is an integrated platform that largely avoids having to fiddle around with MapReduce, Hive, Pig, or any other time consuming and complex technologies. It largely delivers on the promise of big data – ad-hoc access to very large amounts of data, with minimum latency, so that business users can explore, and discover what the data is saying.

The visualisations in Platfora (Vizboards) are rendered in HTML5, and so can be viewed in any browser. However unlike many data visualisation platforms Vizboards can handle millions of data points. Many different types of visualisation are supported, including maps, charts, graphs, dials and tables. The interface is primarily drag and drop with filtering, sorting, grouping and drill down functionality.

Key to the ease of visualisation is the Platfora catalog. It is here that data resources are listed and documented, making it straight forward for users to access and understand the data they need.

The architecture of Platfora is refreshingly straight forward. As can be seen in the diagram below it primarily consists of just four layers – the Hadoop HDFS platform for data storage, the Spark based

data preparation layer, the in-memory architecture for creating data marts (or lenses as Platfora calls them), and finally the data visualisation platform. A catalog provides a semantic view of information, and an API allows external access.

Deployment can be cloud based or on-premises. The Platfora servers are separate to, but live alongside the Hadoop platform. In the cloud, and specifically Amazon Web Services, Platfora uses Amazon Simple Storage Service (S3) to access the raw data and uses Amazon Elastic MapReduce (EMR) to run its data processing jobs (lens builds). The results of the lens build jobs are also written back to S3.

The Lens Builder (re data mart builder) sits over Hadoop and translates requests to prepare data for analysis from the Platfora application into a series of custom Spark and MapReduce jobs. These are submitted to the YARN Resource Manager or Hadoop Job Tracker for execution. Once the data is extracted and transformed within Hadoop, the job results are written back to the Hadoop file system as Platfora lens. The lens is a Platfora columnar file format for storing high performance in-memory extracts of data with analysis applied.

There is much more to the Platfora architecture, but this is probably enough for a short review. It is a highly innovative solution to the decades old problem of business users getting access to data, which in its raw form is not particularly suited to analysis.

Tamr

Tamr employs machine learning technologies to carry out much of the data preparation grunt work. The algorithms work alongside relevant domain experts so that ambiguities and other issues can be resolved. At the simplest level Tamr allows users to register data sources in a centralised catalog, facilitates the creation of a unified schema, cleanses data, and publishes via a RESTful interface a single version of the truth.

Tamr supports a centralised inventory of enterprise data. It automatically catalogs all metadata available to the enterprise in a central, platform-neutral place. This enables data to be grouped by logical entities (customers, partners, employees) rather than where it's stored, making it easier for companies to discover and uncover the data necessary to answer critical business questions.

It allows easy data connections across siloed people, processes and places. Advanced algorithms automatically connect the vast majority of data sources while resolving duplications, errors and inconsistencies among attributes and records. When the system can't resolve connections automatically, it calls for human expert guidance, using people in the organisation familiar with the data to weigh in on the mapping and improve its quality and integrity. Tamr automatically matches attributes across a full range of data sources, often accomplishing up to 90% of the task without human intervention.

Tamr supports structured and unstructured data. This includes CSV and XLSX files, relational databases (via JDBC), semi-structured data, such as JSON, XML and YAML, RDF and full text (via

preprocessing into RDF or semi-structured data). Content management systems such as Documentum, Sharepoint and Alfresco are supported, and of course big data stores in the form of HDFS / Hive, Amazon S3/RedShift and Google Cloud Storage/BigQuery and others.

Tamr presents data, along with its metadata and the results of actions taken in Tamr, in a data inventory. Having the data in one place, along with analysis of the data semantics and connection to other data, makes it easy for users to explore and utilize data that they might otherwise have missed. This data inventory is also integrated with a directory of experts, making it easy to find people within the organisation who are able to answer questions about the data.

RESTful APIs and a variety of export formats makes it easy to tie Tamr into existing infrastructure, allowing data scientists and business analysts to use familiar software, such as QlikTech, Tableau, SAS, IBM Cognos, Recorded Future, Statwing and Zoomdata.

A number of solutions are offered including procurement analytics, clinical trials (CDISC) and customer data integration.

Trifacta

Trifacta provides a self-service data preparation platform that automates many data preparation tasks and allows users to interrogate their data in an efficient manner. The platform learns as users refine their data so that subsequent operations become more automated. It is a highly visual platform with copious graphical representations of data to aid the data wrangling process.

Machine learning algorithms sit at the heart of the capability typically providing rank listed suggested operations that are relevant to the data. A complete data preparation task is stored as a script which can be compiled to run on a variety of systems.

This is a very sophisticated product that will be of interest to large organisations struggling to prepare large data sets for analytical purposes.

A broad range of functions include data assessment, shaping, enrichment, transformation and others, all within the framework of a well governed environment. Assessment gives a high level overview of data quality with detection of missing and unusual values, gaps, data skew, and automatic detection of data types. Enrichment involves the combining of data from different sources to complete the data picture. This might involve using various dictionaries, joining data, creating derived fields and aggregating. Useful transformations can be saved and shared with others in the organisation as a reusable script. Shaping concerns itself with creating generating data at the right level of granularity, and Trifacta uses data inference techniques to introspect the data and automatically apply initial shaping and metadata recommendations for the user.

Trifacta calls its approach to data preparation Predictive Interaction. This leverages a two way interaction between users and Trifacta platform where Trifacta will recommend and users can accept or modify the recommendations. The output from this interaction is a script, and the whole users

interface is designed to provide menus and drag and drop so that the actual act of coding can be avoided.

Trifacta is composed of three integrated layers. Direct data manipulation allows users to select and modify data as needed. The recommendations are generated by machine learning algorithms that learn as the platform is used. These are rank listed suggested transforms that are relevant to the task in hand. All transformations can be viewed in real time in the actual data itself.

The Learning Layer is the province of the machine learning algorithms. These immediately process data and transform it into a usable format as soon as a data source is connected. This includes delimiting, identification of data types (a url for example), and attribute properties with an initial assessment of the statistical distribution of values. The Data Layer supports most data types – from CSV to Hadoop installations.

About Butler Analytics

Butler Analytics is a boutique IT analyst firm specialising in business analytics technologies and methods. It was founded by Martin Butler, best known as founder of Butler Group which, prior to its acquisition, was Europe's largest indigenous IT analyst firm.

Business Intelligence, predictive analytics, big data, fast data, Enterprise Decision Management and all other technologies which aid business decision making are covered.

www.butleranalytics.com
info@butleranalytics.com