Easing the Data Preparation Challenge

Get beyond the hype to make data the lifeblood of your organization

Quest® Software

INTRODUCTION

Understanding data growth

Data growth has skyrocketed over the last few years. In fact, according to research by IDC, the digital universe is doubling in size every two years, and by 2020, we will create and copy 44 zettabytes (44 trillion gigabytes) — nearly as many digital bits as there are stars in the universe.

Driving that data growth is a profound change in how data is produced and what it consists of. Not long ago, most data was generated by systems and stored in databases as structured data. That data had steady, predictable growth — if you were selling 100,000 items, for instance, you would have 100,000 rows in a database table. And the only people interested in that data were IT and financial staff, who had the skills to work directly with the databases.

Today, however, most content is created outside of databases and it takes a wide variety of forms. Some are structured but many are unstructured — people are creating documents, snapping photos and recording videos at a staggering rate. Add in social and sensor data, and the growth rate goes from fast to exponential (see Figure 1).

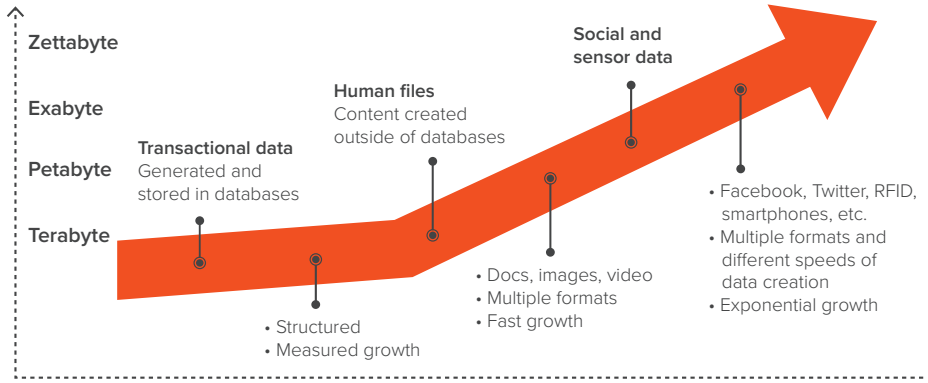

Figure 1. Data growth is now exponential, and much of that data is unstructured.

The success of Big Data initiatives is strongly linked to collaboration between business and IT.

Reaping business value from Big Data

Enterprises recognize that there is a great deal of value in this data — potential insights that can improve decision-making, enhance customer satisfaction, increase employee productivity and drive growth by enabling the organization to capitalize on opportunities. Therefore, a far wider range of people now want to work with the data on a daily basis, many of whom lack the training to work with the databases directly.

How can you enable your organization to reap business value from its data? The hype around Big Data promises that if your organization simply collects enough data, you should be able to answer any question, solve any problem. The truth, of course, is not so simple. Unlocking the value of Big Data requires getting beyond the hype and making accurate data readily accessible to the line-of-business (LOB) employees who need it to speed decision-making.

This white paper explains the barriers to success with Big Data and the best practices that can help you overcome those challenges. One key challenge is data preparation — integrating or blending data from multiple disparate sources to make it available and useful to the enterprise. We explore the five laws of data preparation and show the problems with the traditional data preparation workflow. Finally, we offer a better approach: our Big Data solutions portfolio, including Toad® Data Point.

BARRIERS TO SUCCESS WITH BIG DATA — AND BEST PRACTICES FOR OVERCOMING THEM

You can make data the lifeblood of your business. The first step is to understand the three main obstacles keeping organizations like yours from turning data into insights to drive better decision-making, and the best practices that can help you overcome them:

- Lack of IT and business alignment
- Resource constraints
- Siloed data environments

Lack of IT and business alignment

The success of Big Data initiatives is strongly linked to collaboration between business and IT. Without clearly defined business goals and metrics, Big Data initiatives can get derailed or fail to achieve ROI. Simply put, Big Data projects cannot be bottom-up, IT intellectual innovation exercises.

Accordingly, best practices include the following:

- **Align your project to a business goal** — By clearly articulating the business results you are trying to achieve, you can ensure that the project delivers business value.
- **Measure progress against goals** — Metrics enable you to demonstrate progress to stakeholders.

Resource constraints

High upfront or operating costs can hinder scalability and drain resources. You need to implement effective data management technologies and practices that can put data in the right place at the right time in a cost-effective and scalable way, without rip and replace.

To that end, you should:

- **Leverage existing assets —** By making the most of your existing technology investments, you can reduce upfront costs and drive ROI.
- **Optimize performance —** Ensuring strong database performance will drive user adoption and therefore ROI.

Siloed data environments

To ensure timely access to the data they need, departments often build their own databases or pseudo data warehouses. But having data spread across different systems that are not integrated — often in both on-premises and cloud sources — limits the organization's ability to get the data ready for analysis to enable strategic decisions. Moreover, access to these systems is often limited to a few individuals, keeping business teams from having fast, secure access to the data and analytics tools they need.

Accordingly, you need to:

- **Break down data silos —** You need to understand where the data is, where it is coming from, what its refresh rate is and how it is used. Only then can you blend the data together into an enterprise data warehouse (EDW) to deliver a 360-degree view of all of your enterprise data.
- **Deliver secure, self-service access to LOB staff —** Ensure that the people who understand the business goals have ready access to the data, without assistance from IT.

ACHIEVING HIGH-LEVEL ANALYTICS

The next step is to assess and improve your organization's analytics maturity level, as illustrated in Figure 2. Organizations at the base level can collect data and perform basic analysis, often with simple tools like Excel spreadsheets, in order to answer questions such as, "Who are our top customers?"

By integrating and consolidating data, the organization can move up to the second level and address broader issues, such as comparing the performance of different sales regions. At the third level, adding business intelligence (BI) reporting and analysis tools enables you to start applying key performance

> Reaching the upper levels of the analytics maturity pyramid requires a 360-degree view of your data.

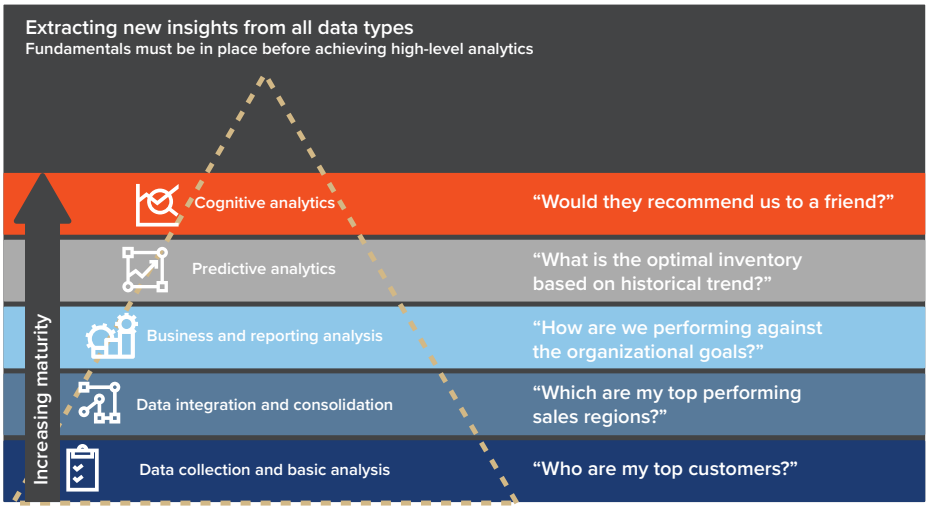

Figure 2. The analytics maturity pyramid

Quest

indicators (KPIs) and assess performance against goals.

Note that at all three of these levels, organizations are working with historical analysis only. To get insights that will enable your company to adapt and change, you need to move up to predictive analytics. By applying complex mathematical algorithms against your data, you can begin to predict what is going to happen in one month, in six months, in 12 months. For example, a manufacturer at the second-highest level could use historical trend data to guide purchasing of supplies and thereby avoid over- or under-buying.

Organizations at the highest maturity level add in cognitive analytics — such as by interrogating data from Twitter feeds, Facebook and other social sources to understand how they are performing against their competitors. For example, by monitoring social media, airlines have dramatically improved their response time to customers who use a social method to contact them, enhancing customer satisfaction and therefore the bottom line.

Note that each level builds upon the ones below it; fundamentals like casual data access need to be in place before you can move to proactive reporting. Similarly, you need to have established reporting systems before getting into predictive and cognitive analytics.

Advancing up the analytics maturity pyramid is necessary because trust is built along this path. In order to accept predictions and other sophisticated analytics, stakeholders need confidence that data accurately represents what has been happening and what is currently happening in the business, as measured by regular KPIs.

THE FIVE LAWS OF DATA PREPARATION

Reaching the upper levels of the pyramid requires a 360-degree view of your data — data from multiple disparate sources must be integrated or blended into a holistic solution. Understanding the five laws of data preparation is critical to the success of your data preparation project.

Law #1: The whole is greater than the sum of its parts.

The notion of process decomposition is deeply ingrained in most analysis techniques used in modern software development lifecycle methodologies. It is based on the presumption that there are natural boundaries along which to divide a complex system into smaller components for integration. This approach comes from the reductionist perspective, dealing with one dimension of problem analysis.

While this approach helps organizations tackle relatively simple problems in short timeframes, it fails as system complexity increases and natural boundaries disappear. All of the gains achieved by breaking down the big problem are lost as the cost of integrating the small solutions becomes unworkable.

For example, consider an automobile. Clearly, creating a working vehicle requires more than simply studying its various parts. In fact, the essence of any end-to-end system simply cannot be captured by studying its individual components alone. Most methodologies fail to realize this fundamental law.

The essence of any end-to-end system simply cannot be captured by studying its individual components alone.

The need for accountability is a critical bylaw. For a data preparation project to succeed, the organization must assign responsibility for the holistic solution — and not only for its initial construction, but also for sustaining it on an ongoing basis.

Law #2: There is no end state.

Organizational entities split, merge and morph into new structures. Political motivations and boundaries change. Technology evolves, and what's leading-edge today is legacy tomorrow. Therefore, an effective data preparation project must consider the full lifecycle of a system and follow best practices that recognize the adaptive nature of complex systems. From the start, we must plan for constant change.

In addition, while Big Data hype tends to focus on new data from the latest technologies, such as social media and the Internet of Things (IoT), organizations must also integrate data from their legacy systems. There have been many waves of technology over the years that seem to move in regular seven-year cycles (such as mainframe to mini to micro computers, and monolithic to client/server to web-service applications). The shift from one wave to the next is never instantaneous; in fact, a given technology will usually last through several waves before it is fully replaced. Therefore, the data preparation project must deal with three or four generations of legacy technology simultaneously.

This is not to say that everything you see should be included in your solution. In particular, it is a natural human tendency to be enamored with new things. But just because something is shiny and new doesn't mean that it fits into your data preparation project — it must be justified as providing value. The data preparation project must guard against adding new systems, data and functionality in a piecemeal fashion, either organically, through acquisition or otherwise. Instead, you must constantly balance the drive for delivering new functionality quickly with the discipline to rationalize the legacy environment to keep up with the changes.

Law #3: There are no universal standards.

There are many different standards out there: Oracle, SQL Server, Hadoop, HDFS, MongoDB, DynamoDB, SimpleDB and many more. Even successful standards (such as TCP/IP for the internet) are not universal. And when it comes to software standards such as COBOL or Java, interoperability and transportability come at the expense of vendor-specific extensions, forcing developers to use a less-than-ideal core set of "pure" language features.

You need to integrate data that follows many standards in order to get a 360-degree view of your enterprise data. You should strive to define and adopt standards within the enterprise, and also work externally with data suppliers and standards organizations to gain agreement across the industry. That said, you must deal with the reality that many forces — including competition, the "not invented here" syndrome and evolving technologies — will result in many different standards for the foreseeable future. Understanding this law will give you the proper mindset: that data preparation is going to be a challenge and not just a simple project.

Law #4: Information adapts to meet local needs.

The Information Engineering movement of the early 1990s was based on the incorrect notion that an enterprise can have a single consistent data model without redundancy. A more accurate way to look at information is as follows:

Information = Data + Context

This formula says that the same data can have different meanings across different domains. For example, a simple attribute such as "Current Customer" can mean different things to the marketing, customer service, and legal departments. Therefore, you can't assume that "Current Customer" in a table from one data source means the same thing as "Current Customer" in a table from another data source. In fact, the field might be an integer in one database and a text field in another.

Another example might be the attribute "Gender." In many tables, this has only two states, male or female, but newer data sources may recognize many more values for "Gender." On a more fundamental level, a given word can take on different meanings in different communities — consider "Date" in a sales database versus an online dating site. The data preparation project must embrace informational diversity and recognize that variations exist, and use techniques to compensate for them. In particular, it is imperative to include subject matter experts (SMEs) on the project team.

Just because something is shiny and new doesn't mean that it fits into your data preparation project — it must be justified as providing value.

Data profiling helps you not only to understand anomalies and assess data quality, but also to discover, register and assess enterprise metadata.

Law #5: All details are relevant.

Albert Einstein famously said that problems cannot be solved on the same level of thinking that created them. Solving a problem requires abstraction. But abstraction is the practice of representing a problem without all the details, developing a model based on that representation, and then using the model to create the real-life solution. Therefore, the effectiveness of an abstract model is inversely proportional to the complexity of the context — as you abstract away more details, you risk losing something important. No detail can be safely ignored. Moreover, the cost of developing and maintaining abstract models of the system can become an economic black hole, consuming all benefits.

There's no question that Big Data projects are going to have a huge impact in the coming years. However, we should not be throwing away all legacy technology and legacy data. Instead, we should be looking at how to integrate legacy data and Big Data using an abstract model that has room for both. Otherwise, we're not going to get the result that we want and need.

THE TRADITIONAL DATA PREPARATION WORKFLOW

For many organizations, the data preparation workflow proceeds as illustrated in Figure 3. Once you connect to a data source, the first step is to profile the data — analyze its structure, content, relationships and derivation rules. Profiling helps you not only to understand anomalies and assess data quality, but also to discover, register and assess enterprise metadata. Data profiling is a critical input task to any database initiative that incorporates source data from external systems or more than one internal system.

If you have just one database or EDW, you may be using the profiling tool that came with it. If you are dealing with multiple different data sources, you may be using a third-party profiling tool like Datamartist, Ataccama, Grid-Tools, Talend, Informatica or DataCleaner.

Once you have a data profile from the first tool, you use a second tool to report on that profile. Then you use a third tool to alter the data query, and then you repeat the cycle until you understand the data. It is critical to recognize that not everything can be profiled quickly; it can take time to get to the point where you understand the data.

Eventually, you can use yet another tool to push the data out to your LOB associates so they can analyze it and use it to drive decision-making.

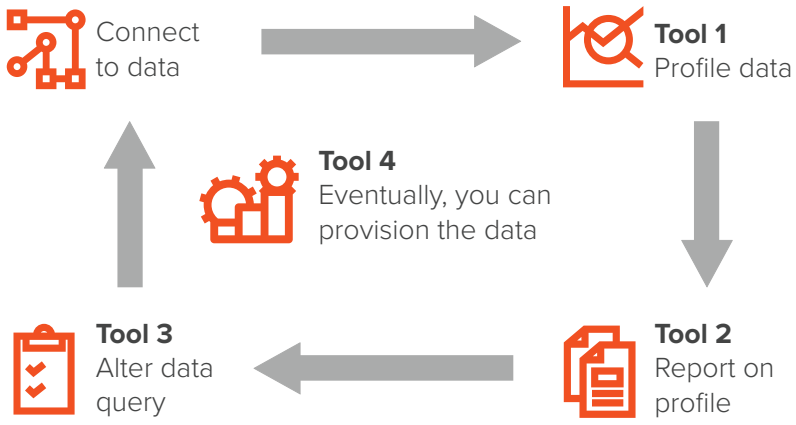

Figure 3. The traditional data preparation workflow

HOW QUEST® CAN HELP

Toad® Data Point

Toad Data Point enables you to do all of these tasks with a single tool. With Toad Data Point, you can easily connect to nearly any data source; quickly develop queries to profile, transform and cleanse the data; and then provision it — all from one interface (see Figure 4).

Just what do we mean by "nearly any" data source? Here's a partial list:

- **RDBMS sources** like Oracle, SQL Server, SAP, DB2, Teradata, MySQL, Access and any ODBC-compliant platform
- **NoSQL sources** like Salesforce.com, Hadoop, Cassandra, MongoDB, SimpleDB and DynamoDB
- **BI sources** like OBIEE, Business Objects and Microsoft Analysis Services

Even better, Toad Data Point puts all this power into the hands of everyone in the enterprise who needs to work with your data. Its easy-to-use drag-and-drop interface eliminates the need for SQL expertise. You can even use a local data store as a sandbox to review, manipulate and profile data prior to analysis. This allows you to work with data offline, without impacting production systems. And you can automate and schedule routine tasks so you can be more productive.

An end-to-end Big Data solutions portfolio

Toad Data Point is just the start. Quest offers an end-to-end Big Data solutions portfolio that is platform-agnostic and data-agnostic to enable your success now and into the future. We have combined new product development with an acquisition strategy and services to deliver modern, comprehensive solutions that fill the gaps created by fragmented niche products while avoiding the lock-in penalties of rigid and costly legacy solutions (see Figure 5).

Our portfolio includes:

- **Database management tools** such as Toad for Oracle and Toad for SQL Server
- **Data integration tools** such as SharePlex®
- **BI, data access and data discovery** tools such as Toad Data Point and Toad Intelligence Central
- **Integrative solutions** that work well with your other toolsets such as the Statistica platform and Dell Boomi

While some of these capabilities are relatively new under the Quest umbrella, these solutions have proven to be industry-leading for more than 30 years. Plus, our service offerings ensure we'll help you get above the hype to focus on what Big Data can do for your organization.

Toad Data Point's easy-to-use drag-and-drop interface eliminates the need for SQL expertise.

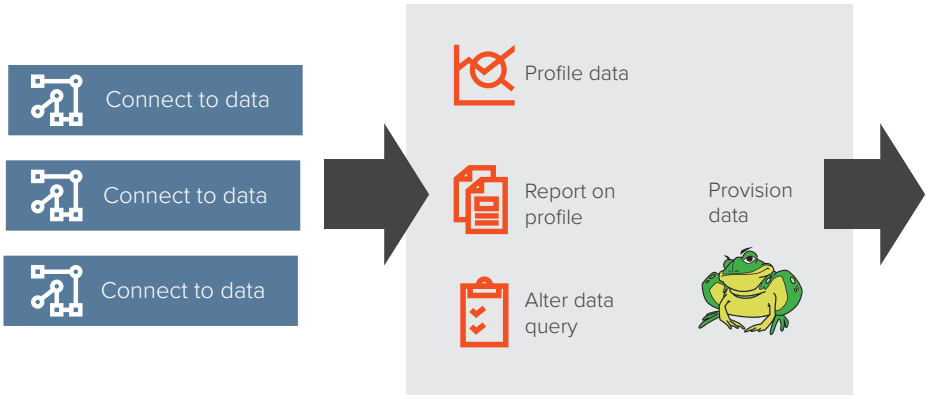

Figure 4. Toad Data Point delivers the complete data preparation workflow in a single tool.

Quest

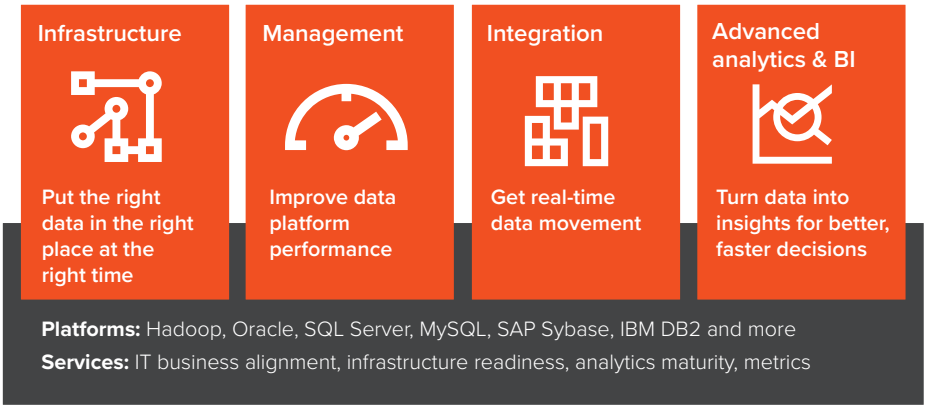

Figure 5. Quest has a comprehensive yet modular solution for analysis-ready data.

By ignoring the hype and focusing on business objectives, you can achieve your Big Data goals.

Quest solutions are modular, meaning we don't impose our view on any one portion of the Big Data stack, so you're able to choose where to start and how to prioritize your project. You can work with Quest without worrying about starting over, and we'll empower you to maximize your potential to turn data into insights for better, faster decision-making.

CONCLUSION

Follow these easy guidelines to ensure quality analytics in your organization:

- Build a cross-divisional team to keep the data preparation project on track. Ideally, you should include a technical resource who understands how integration tools work, business users who will run analytics based on newly integrated data and create reports, and company decision-makers — the consumers of the information.
- Follow the five laws of data preparation.
- Understand your data sources and data types, and ensure that users can connect to them easily.
- Profile your data before building your data views. You need to understand where the data is coming from, what it looks like, and how it will impact the next data source when you add it together.
- Seek tools that empower business users with visual user interfaces, drag-and-drop templates and automated workflows.
- Check and double check your data before releasing it.

By ignoring the hype and focusing on business objectives, you can achieve your Big Data goals.

ABOUT QUEST

Quest helps our customers reduce tedious administration tasks so they can focus on the innovation necessary for their businesses to grow. Quest® solutions are scalable, affordable and simple-to-use, and they deliver unmatched efficiency and productivity. Combined with Quest's invitation to the global community to be a part of its innovation, as well as our firm commitment to ensuring customer satisfaction, Quest will continue to accelerate the delivery of the most comprehensive solutions for Azure cloud management, SaaS, security, workforce mobility and data-driven insight.

If you have any questions regarding your potential use of this material, contact:

Quest Software Inc.
Attn: LEGAL Dept
4 Polaris Way
Aliso Viejo, CA 92656

Refer to our Web site (www.quest.com) for regional and international office information.