CLUSTERPOINT

THE SECRET TO EASY, FAST, ECONOMICAL BIG DATA
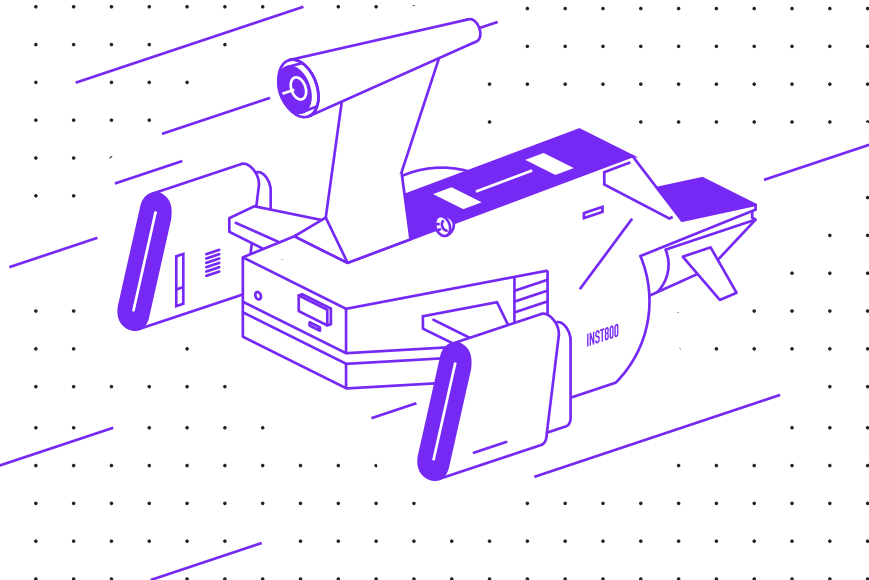
Jason Bloomberg, Intellyx, Forbes
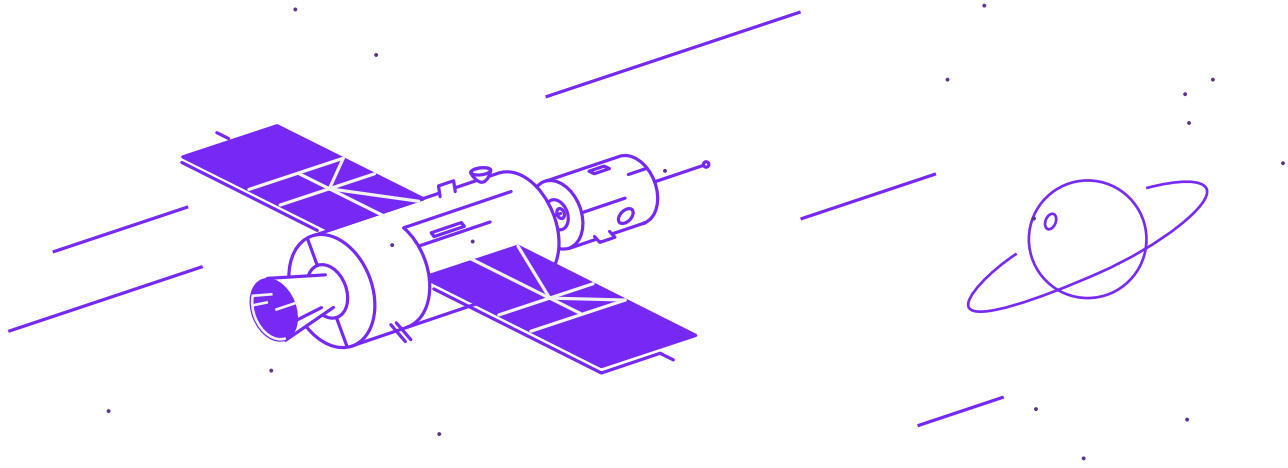
UNDERSTANDING THE STRATEGIC DIFFERENTIATION OF THE CLUSTERPOINT DBAAS

With all the talk today of the three V's of big data – *volume, velocity,* and v*ariety* – few people like to mention the three C's: just how *confusing, complicated,* and *challenging* the big data world has become in the few short years since the buzzword hit the scene.

Even the definition of big data presents a confusing challenge: *data sets too large for conventional tools to store, process, or analyze.* Big data are a moving target by definition, as new tools enter the market, confounding any notion of *conventional.*

Make no mistake – new tools arrive on the scene every day. New databases, data storage technologies, data movement, integration, and messaging products, as well as analytics, visualization, and insight technologies. Even emerging market categories are in flux, from NoSQL ("not only SQL") to DBaaS (Database-as-a-Service) to "data intelligence" (whatever that is).

Out of Latvia comes a recent entrant to this confusing market: *Clusterpoint.* Clusterpoint offers ACID-compliant NoSQL DBaaS with built-in JavaScript querying. It's blisteringly fast, less expensive than other DBaaS alternatives, and its JavaScript capabilities run circles around existing JavaScript-centric document stores that have been on the market for a while now.

But comparing product facts and figures, benchmarks, and metrics isn't the whole story. To understand Clusterpoint's strategic value proposition, we must place its innovations into the broader context of today's disruptive enterprise trends: *the democratization of big data, data gravity,* and the increasingly important story of *real-time big data.*

DEMOCRATIZATION OF BIG DATA

The old days, when IT management fastidiously maintained control and access to technology assets, reluctantly doling them out only when hapless users jumped through enough hoops, are long gone.

Today, everyone carries a supercomputer in their pocket. Customers, partners, and employees at all levels of the organization take access to technology for granted.

The two sides of this democratization trend are *the ease of access and the ease of use* of the technology available to everyone. As technology touchpoints multiplied, so too did the usability and overall customer experience for the applications IT – and other, third party sources – brought to the organization.

The democratization of IT has upended the notion of an enterprise application as well, and digital transformation initiatives drive the assembly of increasingly complex, distributed, mainly cloud-based apps that deliver enterprise value to technology consumers – within companies of all sizes and among the public at large.

The world of big data is also subject to ongoing democratization, as an increasing number of people across the enterprise use analytics and visualization tools to interpret data, gain critical business insights, and make decisions.

But there are roadblocks impeding this progress toward self-service big data. Today, big data are too difficult, too slow, and too expensive. Dealing with massive data sets still presents a challenge, even in more advanced organizations.

To address these challenges. Clusterpoint brings big data processing to a broader business audience – both business users as well as a burgeoning class of digital technologists, many of whom are comfortable with user interface-centric technologies like JavaScript.

In fact, Clusterpoint's JS/SQL combines JavaScript and SQL – two broadly understood, familiar languages that are well within the grasp of digital professionals both inside and outside IT shops. True, no one would expect a CEO or a COO to work with these languages directly, but their broad popularity still frees big data from the confines of the IT organization – an essential part of the democratization story.

Clusterpoint's pricing approach also supports the trend toward democratization and self-service. Unlike other DBaaS offerings on the market today, Clusterpoint offers a flexible pay-per-use model, where people only pay for computational resources they actually use. Furthermore, it offers the first 10 GB of storage free – a low barrier to entry that is plenty for most users to get their projects up and running.

Because Clusterpoint frees users from worrying about scalability, hardware sizing, and the cost of provisioning storage, people at different levels and in different parts of the organization can treat Clusterpoint as a self-service resource – while in fact, it is a powerful, sophisticated database behind the scenes.

DATA GRAVITY: BRING COMPUTE TO THE DATA

Everybody realizes at an intuitive level that it takes a certain amount of time and money to move each datum from *here to there*. The speed of light *always* provides a theoretical maximum velocity, and even at the chip level, moving electrons from CPU to level 2 cache is the fundamental bottleneck for all compute processing.

Now multiply that single datum by many trillions, and this problem of *t* becomes a serious consideration for all big data. When we're dealing with many terabytes, even moving from one instance to another in the same cloud will run up the bill and slow everything down.

> TO ADDRESS BIG DATA'S DATA GRAVITY CHALLENGE, THE RULE OF THUMB IS TO MOVE THE COMPUTE TO THE DATA.

After all, shifting where the processing takes place is straightforward in today's fully virtualized world.

Surprisingly, many big data technologies don't follow this simple rule. For example, MapReduce running on Hadoop follows a "database to query to compute" pattern, where MapReduce jobs perform exhaustive processing on all the data in a pre-populated HDFS cluster.

When organizations are leveraging Hadoop for batch processing, this approach may be adequate – although even with batch, processing times can be excessive. In many situations, however, organizations require a better way of dealing with the data gravity issue.

Clusterpoint solves the problem of data gravity in a novel way by combining two technologies. First, by embedding JavaScript processing directly into SQL statements, they deliver a "database to compute and query" pattern instead of the traditional "database to query to compute" approach. In other words, *compute and query become the same step* – an improvement that becomes increasingly dramatic as data set sizes explode.

Secondly, Clusterpoint uses indices to access data, so exhaustive processing isn't necessary. Clusterpoint's indexing is immediate and automatic, so all information in a Clusterpoint database can benefit. Indexing has long been an important part of big data processing to be sure, but Clusterpoint builds it into every data set behind the scenes.

BIG DATA IN REAL-TIME

A central challenge for any big data initiative is dealing with the ever-increasing velocity – of the data themselves, as well as the increasing velocity of the business. Squeezing every last millisecond of performance leads to the demand for *real-time* – technology with no delays whatsoever, moving at the speed of thought itself.

There are important nuances, however, to this notion of real-time. First, real-time never actually means *instantaneous*, as it always takes a certain amount of time for data to find their way to their destination.

For many applications, the most important definition of real-time is *low latency*. Latency refers to how long a web site or app takes to respond to a click or other user interaction (either on a computer or a mobile device), and thus the faster, the better.

Real-time may also refer to *up-to-date information*. In a breaking news situation, for example, people want the very latest information. Real-time airline or theater seat availability falls under this definition as well.

A third sense of real-time refers to *human interactions*. Gamers playing a multiplayer game, for instance, want the action to be real-time. Online voice conversations and some fast-paced auction sites also require this type of real-time.

Finally, we may be referring to real-time *processing of information*. Stock trading and online ad placement are two of the most familiar examples. Lowering latency is part of this challenge, but the bottleneck isn't just serving up information to the user – it's all the number crunching behind the scenes that must also take place at a blisteringly fast pace.

IN THE CONTEXT OF BIG DATA, THE MOST IMPORTANT SENSE OF REAL-TIME IS OFTEN UP-TO-DATE INFORMATION, WHEN THE GOAL OF THE BIG DATA PROCESSING IS BUSINESS INSIGHT.

Decision makers no longer want to wait until the end of the day or week to have information they can act upon. Instead, they want real-time information they can act upon immediately. Real-time big data processing is also increasingly important for a variety of automation scenarios – essentially, situations where an organization wants to take a human out of the loop. Adjusting stoplight timing to improve traffic flow is one situation where low-latency and real-time information processing become essential tools – and by extension, any other real-time process, from manufacturing to airport flight control.

It's essential to realize, however, that real-time big data processing depends upon the ability of the underlying technology to scale properly. Any bottleneck – in storage, queries, transactions, logic execution, integration, or elsewhere – can slow down the entire application end-to-end.

Cloud-based, elastic horizontal scalability is merely the price of admission today – and clearly, Clusterpoint couldn't actually offer DBaaS unless they had such scalability. But there's more to Clusterpoint's real-time story: indexing and ACID compliance.
Clusterpoint's indexing provides real-time queries over arbitrarily large data sets – and takes indexing one step further by adding the ability to provide logical joins of large documents. As a result, queries take place in real-time even over multiple data sets.

Furthermore, ACID compliance is important to real-time behavior because ACID requires immediate consistency, while many Cloud-based databases only offer eventual consistency. Eventually consistent databases may commit transactions locally in those situations where the nodes are unable to commit transactions globally – thus introducing a delay when an application requires global consistency.

Cloud database aficionados will recognize the challenge of consistency from the CAP theorem. The CAP theorem states that no database management system can offer immediate consistency, partition tolerance, and high availability at the same time.

Clusterpoint takes a quorum-based approach, were most of the nodes must be able to communicate in order to commit a transaction. But since Clusterpoint controls their own deployment environment (running in their own collocated data centers rather than AWS), the chances that such a quorum won't be achievable is minuscule. The end result:

CLUSTERPOINT PROVIDES ACID TRANSACTIONALITY, WHERE THE C STANDS FOR IMMEDIATE CONSISTENCY – IMMEDIATE IN THE SENSE OF REAL-TIME.

THE INTELLYX TAKE

In the enterprise, big data do not stand alone. They are a part of the fabric of the business, as enterprises become software-driven organizations.

It should come as no surprise, therefore, that the broader transformational trends that are upending companies around the world are also transforming how organizations collect, store, process, deliver, and use data.

The three trends this paper calls out – the democratization of IT, data gravity, and real-time behavior – impact the organization much more broadly than a discussion of the transformative impact they have on big data themselves.

There's no question, however, that data are the lifeblood of today's organizations. From healthcare to banking, retail to manufacturing, data are the secret sauce to maintaining a focus on customer needs while remaining agile and innovative.

Clusterpoint is well-positioned to ride this wave. Their savvy use of the Cloud, both to provide seamless scalability as well as cost-competitiveness, combined with their novel embedding of JavaScript into SQL queries, delivered in real-time – makes them a worthy competitor in the crowded and tumultuous database marketplace.