GEDDES Fresh thinking

Slow Data
VERSUS
Quick Data

A WHITE PAPER

THE UNDER-APPRECIATED RISKS OF BIG DATA AND CLOUD COMPUTING

Big Data requires Slow Data and Quick Data, and you need to put these in the right place to increase rewards and reduce risks. CIOs should be concerned about whether they are making the right choices.

Supercomputer designers have long faced similar challenges: whilst the timescales are different, the issues are the same. There is a transferable understanding that can help CIOs to make good decisions.

In this paper we offer guidelines from supercomputing to help CIOs to measure and manage the risks of using cloud services, and increase the potential for rewards.

Measure and manage the risks of using cloud services, and increase the potential for rewards.

Big Data is getting cloudy, and this is not a simple matter

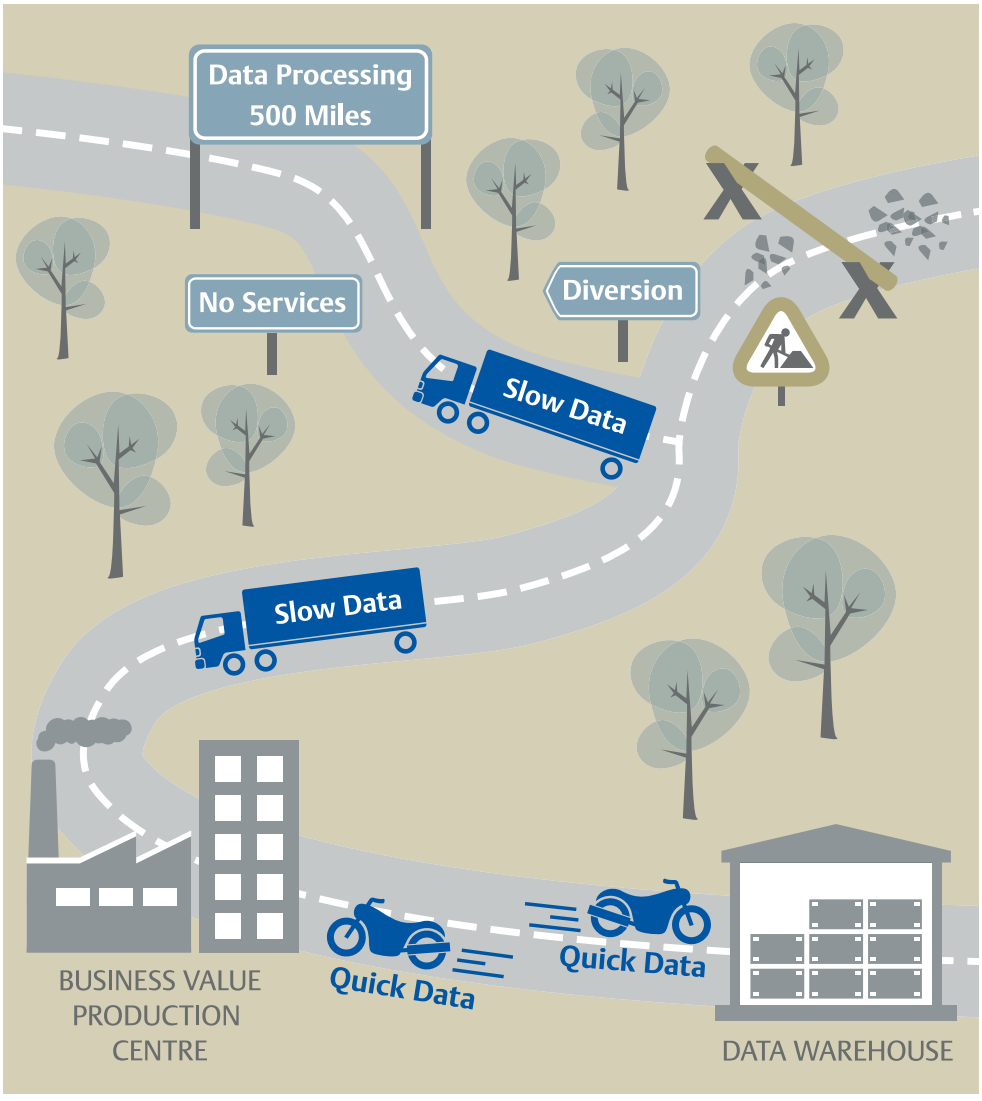
BIG DATA IS RE-IGNITING ECONOMIC GROWTH

Western economies are slowly returning to economic growth after the financial crash of 2008. One basis for this is that enterprises are increasingly driving corporate strategic advantage by using the latent value of data they have collected. This approach requires the capability to sift and organise huge and diverse amounts of data into actionable business information, a process that is enabled by what is colloquially known as Big Data technology platforms.

THE CRITICAL PERFORMANCE PATH

All enterprise data feeds business processes, and every process has a critical path through it. Some data is needed more urgently because it sits on that critical path; delays to its timely arrival at the point of use directly impact business performance. Other data can either take longer to arrive, or there are ways of disguising its latency, such as pre-caching it near where it may be consumed in future. Any failure to understand the location and nature of the critical path by definition creates cost and performance hazards.

Figure 1. **SLOW DATA VS QUICK DATA**



Quick Data is most sensitive to latency; Slow Data is most sensitive to processing cost.

SLOW DATA AND QUICK DATA

As a result of this critical path effect, Big Data assumes two distinct forms: Slow Data and Quick Data. Slow Data is off the critical path; Quick Data lies on it. Furthermore, the need for timely action is constantly nudging the boundary between these. The need to create a strategic business advantage or respond to competition means that Slow Data can become Quick Data. Meanwhile, formerly Quick Data can become Slow Data as automation shifts critical paths. Quick Data is most sensitive to latency; Slow Data is most sensitive to processing cost.

INCREASING LEVELS OF VIRTUALISATION

The growing volume and diversity of demand for data processing, coupled with a drive to reduce cost, is prompting a qualitative change in the technical infrastructure used to process this data.

Enterprises have already largely moved away from applications having their own dedicated servers. They instead use virtualised applications running on shared (but still private) on-premises enterprise infrastructure. This increases resource usage efficiency, whilst applications remain accessible over a local area network (LAN).

The next step, and one that many are taking, is moving to a centralised and shared private virtualised infrastructure, or to public cloud platforms. In this case, applications are accessed over a wide area network (WAN).

THE TRADE-OFF

As a result of these changes, we now have a chain of virtualised and shared physical resources – CPU, transmission, I/O – with many applications contending to access them. Each evolutionary architectural change raises resource usage efficiency and allows for greater peaks in demand.

However, this resource-usage efficiency comes at a price. Cloud distributes data away from points of creation and consumption, and virtualisation creates queues of work waiting for resources.

Thus both virtualisation and WAN access alter the distribution of transaction times: it increases both their average latency and their variability. This effect is magnified as more round trips are required.

Hence there is a trade-off: whilst Slow Data benefits from virtualisation by gaining cost efficiency, all data access takes a performance hit; and Quick Data takes a hit that creates new application performance and business hazards.

Whilst Slow Data benefits from virtualisation by gaining cost efficiency, all data access takes a performance hit; and Quick Data takes a hit that creates new application performance and business hazards.

The paradox of distributed (cloud) computing is that sometimes it involves increased centralisation of computing, in order to get the right balance of cost and performance.

THE NETWORK IS THE MAIN CONSTRAINT

These performance problems arise from many root causes. To correctly determine which is the source of a specific problem requires a robust model of cause and effect. Frequently correlation is taken as proof of causation, but the strongly coupled nature of these complex systems means this is often a misguided and costly assumption.

Typically it is the network that is the system's performance constraint. You can't undo delay or reverse time, so you can't hide the latency for Quick Data on the critical path. The system is only as good as its weakest link, and latency caused by network-induced delay is that weak point.

Since the speed of light is not negotiable, that means the proximity of Big Data to its points of creation and consumption matters greatly. The paradox of distributed (cloud) computing is that sometimes it involves increased centralisation of computing, in order to get the right balance of cost and performance.

THE CIO'S NIGGLE

CIOs are aware of these issues, and thus walk around with a 'niggle': am I making the right trades? Specifically, am I getting the upside: significant cost improvements and value from flexibility? And am I managing the downside: have I really covered my risks, and will they mature into failures on my watch?

There are good reasons to believe many should be paying more attention to that niggle.

These questions are the very same ones that supercomputer designers face. How do I make the trade-offs between where to compute and when to communicate, in order to get the optimum cost, performance and reliability? Their problems and principles are therefore worthy of our attention.

Quick Data and Slow Data have differing architectural needs

The diverse demands placed by the different types of data and business process create a requirement for an equal diversity in supply. Regrettably, this is often overlooked. There are six specific problems that result.

PROBLEM #1 – BAD EXPERIENCES BECOME MORE COMMON

Business processes fail due to outliers of transaction time. These bad experiences drive customer frustration and then churn. That means you have to make bad experiences sufficiently rare, rather than simply making good experiences relatively common.

Going to the cloud changes the distribution of transaction times and tends to increase the size of the 'tail of failure'. Quick Data that used to arrive 'just in time' now becomes 'just too late'.

However, business process management may not be measuring the 'tail' at all, but merely measuring average transaction time. Even if transaction latency is measured, there may not be means to manage and mitigate performance problems in the 'tail'.

Going to the cloud changes the distribution of transaction times and tends to increase the size of the 'tail of failure'. Quick Data that used to arrive 'just in time' now becomes 'just too late'.

PROBLEM #2 – CATASTROPHIC FAILURES BECOME MORE LIKELY

Cloud computing platforms have complex non-linear performance behaviours that are not captured by simple modelling. These behaviours are akin to 'weather' and 'climate' in the physical world. Small failures can condense and coalesce to create storms. Moving to the cloud increases the chance of 'extreme weather events' that create major business outages, of the kind that create PR crises.

The early warnings need to be measured and the risks managed, rather as 'amber' and 'red' warnings alert us to impending severe weather. However, cloud operators and their customers often understand the risks only poorly. Though the risks may remain hidden, ignorance will not be bliss as these hazards mature into service failures.

Moving to the cloud increases the chance of 'extreme weather events' that create major business outages, of the kind that create PR crises.

When data finds itself in the wrong place for processing, there is a need to buy expensive long-haul capacity to relocate or replicate it.

PROBLEM #3 – LACK OF WARNING OF PROBLEMS

Large, complex distributed systems also exhibit non-linear performance characteristics. When these performance hazards are unknown, then the enterprise faces unquantified contingent liabilities. There are consequent (uncosted) risks in mitigating them. In particular, when data finds itself in the wrong place for processing, there is a need to buy expensive long-haul capacity to relocate or replicate it. For Quick Data, there is an additional cost overhead in keeping latency low, which may require high capacity links with very low utilisation.

PROBLEM #4 – MISATTRIBUTION OF CAUSE AND MISALLOCATION OF COST

IT organisations lack adequate tools to identify which data is Quick Data and which is Slow Data. Existing techniques thus fail to reveal the contribution of each component and data source to the overall system performance. As a result, CIOs will increasingly find themselves moving data sets and functionality around on the basis of weak correlation effects rather than strong models of causation. If you can't tell Quick Data from Slow Data, you will end up having to treat all data as Quick Data with the attendant high networking costs.

If you can't tell Quick Data from Slow Data, you will end up having to treat all data as Quick Data with the attendant high networking costs.

PROBLEM #5 – FALSE SENSE OF SECURITY

There is a general assumption that problems can be fixed later, as and when they occur, just by throwing bandwidth at them. There is an underlying belief that any performance issues can be mitigated by a tactical spend to increase capacity, and that the costs of doing so will be modest.

Too often, this philosophy doesn't work. You can't use faster link speeds to overcome speed of light. The end result therefore tends to be unplanned application performance declines and/or very high communications cost growth. Quick Data becomes Never Quick Enough Data. Can you completely re-architect your cloud solution, and keep your business running, when performance problems arise from having your data in the wrong place?

Can you completely re-architect your cloud solution, and keep your business running, when performance problems arise from having your data in the wrong place?

Outsourcing an infeasible requirement does not take away the risk, it just grows and defers it.

PROBLEM #6 – LOSS OF CONTROL OVER FIXING THE PROBLEM

The CIO has been tasked by the CFO and CEO with cost-optimising the data processing infrastructure. However, cloud is a different space, and is a fundamental change. On the enterprise's LAN the choices over performance had always been relatively constrained, but the CIO had complete control. Now the design choices have become far more complex, and the scope for control is limited. Once you move to the cloud, there is no way back, since the old infrastructure has usually been de-commissioned.

CIOs may have created service-level agreements for all the components in isolation, but the performance risk for the overall system integration always resides with the enterprise. No SLA clause can compensate for a catastrophic business outage; at that magnitude of failure, your supplier is potentially out of business too. Hence outsourcing an infeasible requirement does not take away the risk, it just grows and defers it. That leaves CIOs accountable for a risk over which they have insufficient visibility and control.

Cloud is a risk that could kill your business

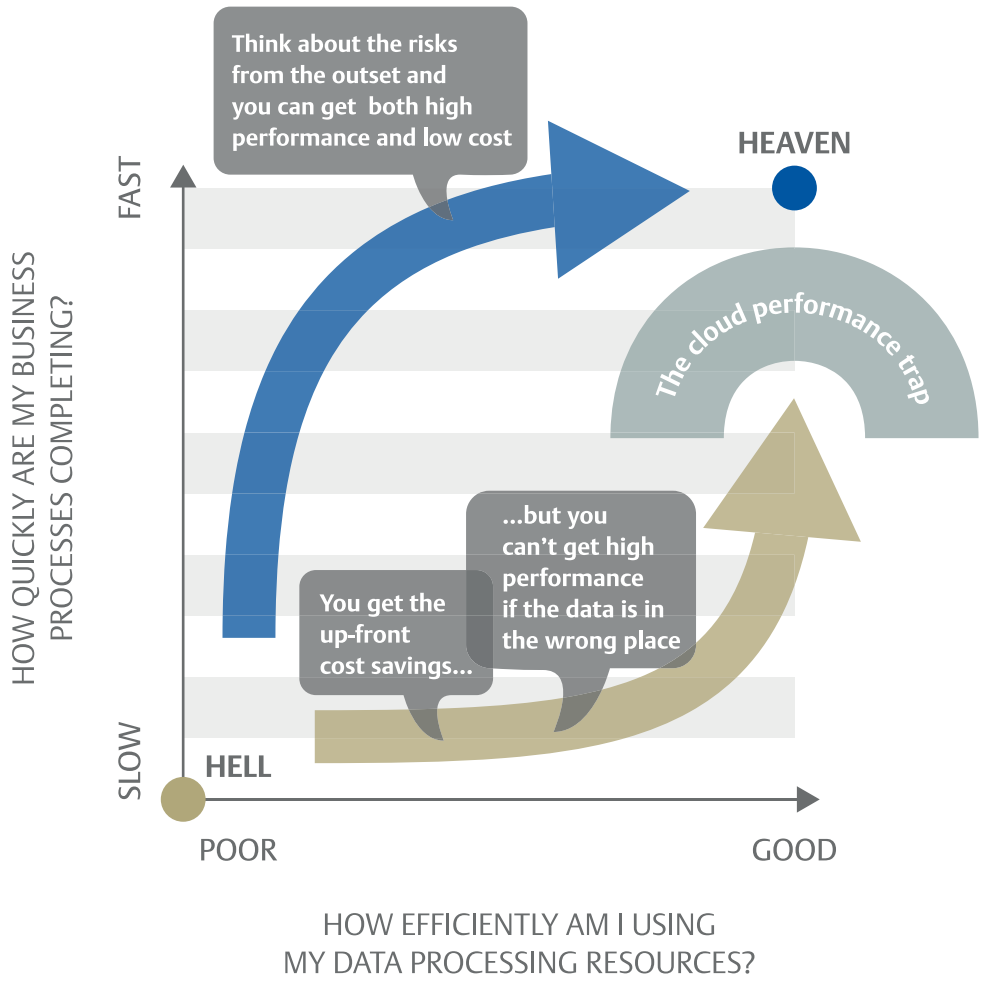
There are inherent risks in moving to the cloud: new distributed computing design and performance and issues; limited operational control over network and computing resources; and increased reliance on third-party suppliers for core business processes. When you multiply these hazards by their potential impact, there is a lot of (often hidden) risk.

These challenges provoke the need for a different and richer kind of conversation between enterprises and their cloud suppliers.

While every vendor will tell you that their platform, network or application is the best one, what happens when you put all the components together? Will the result be fit-for-purpose and deliver the right combination of cost, quality and risk? Does the vendor understand Failure Modes Effect Analysis, and can they deliver credible quantitative models to support their design decisions?

Failure to address these issues causes a lack of alignment of concerns along the supply chain. Your objective may be low latency to enable a high completion rate for a business process. The supplier's goal may be high server uptime and intense resource utilisation. The disconnect means that the promised rewards cannot be realised.

Figure 2. FLOW EFFICIENCY



The danger lurking ahead for CIOs is that having selected the best-of-breed technology and service vendors, they feel they can sit back and relax. Yet after building the system, the CIO may find they get neither the resource usage efficiency nor high performance nor benefits of Big Data. Slow Data has Quick Data costs; Quick Data has Slow Data performance. A few years down the road, small fires break out; CIO time is increasingly dedicated to fire-fighting performance issues, rather than creating strategic advantage through Big Data.

One day, there is an upturn in business combined with sudden performance problem. There is a cataclysmic systems failure, which takes weeks to recover from. You go out of business, as you can't get out of the vicious cycle of making bad strategic trade-offs and pouring resources into tactical performance fixes.

If instead you make good trade-offs you can get both resource usage efficiency and performance. To achieve this you have to know where do you put stuff that is urgent to you, and be able to decide if the cost saving of shared cloud resources is worth the money gain.

How to put the right data in the right place in the right kind of home? Strategic business advantage comes from knowing how to make these trades.

How to make the right trades and put things in the right location

Where to locate data stores and computing, and how to make the trades, is a solved problem in the domain of supercomputing. Here are some questions that we suggest would make an excellent opening for a conversation between any CIO and their cloud platform supplier.

1. **Technology:** Do you understand and know how to apply the core techniques from supercomputing to cloud architecture and operation? Are you aware of the existence and nature of the trades, and in particular pay attention to the performance hazards that exist?

2. **Organisation:** Who is the Performance & Risk Tsar responsible for overall system and business outcomes? What new organisational capabilities does cloud computing require? How do you get visibility of the trades being made, which are often implicit? Who makes the calculation of the right trades? Do you have the skills required to pick the right locations for storage and computation? Which technical staff have experience in supercomputing, parallel processing or safety-critical systems design?

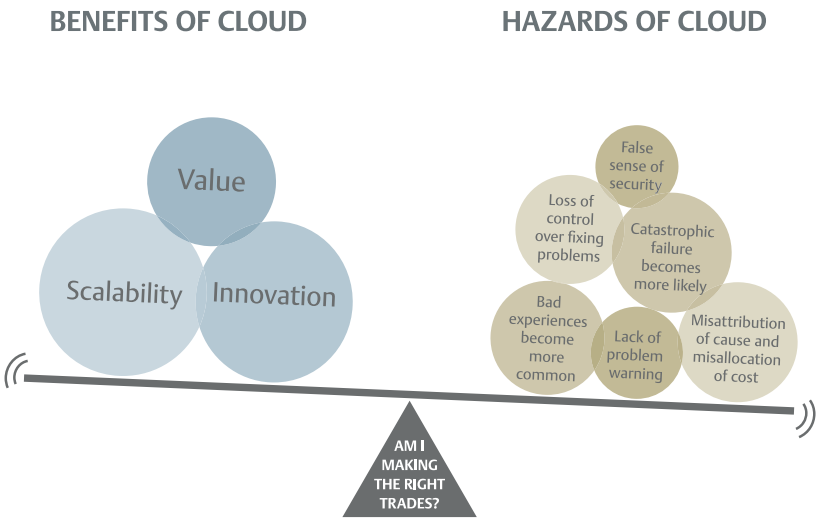
3. **Data modelling:** How can we perform the critical path analysis to understand which data is Slow Data vs Quick Data? How can we separate the 80% Slow Data from the 20% Quick Data to create performance management processes appropriate to their different needs? How can their changing needs be accommodated? Do I have an architecture that permits me to migrate Quick Data to Slow Data, and vice versa?

4. **Operational process:** What is the best way to 'Measure it; model it; mitigate it'? What are the appropriate system performance measurements and metrics? What management controls are needed to review and revise the location of each data and compute function?

5. **Design for performance:** Should you put your Quick Data in low-latency co-located hosting, close to other clouds? Is data in the right location as more complex industry supply chains form? When does "best connected" beat "lowest price"? When should you measure by the cost of milliseconds, not that of microprocessors?

6. **Design for efficiency:** When should you locate Slow Data in distributed or lower-cost environments (e.g. Amazon or Google)? When is it best to co-locate Slow Data with Quick Data for economies of scale and performance flexibility? How can you hedge your risks of Slow Data becoming Quick Data?

Figure 3. TRADE-OFF



Contact Information

WE CAN HELP

If you would like to discuss these questions with us, or would like us to identify and quantify those most pertinent to you and your organization's information processing needs, please do get in touch. We can help you to measure, model and manage the risks of using cloud services, and increase the potential for rewards.

Register for the free *Future of Communications* newsletter at www.martingeddes.com

Registered Office	Martin Geddes Consulting Ltd 17-19 East London Street Edinburgh EH7 4BN United Kingdom
Email	sales@martingeddes.com
Web	www.martingeddes.com
Linkedin	www.linkedin.com/company/990245
Twitter	@martingeddes