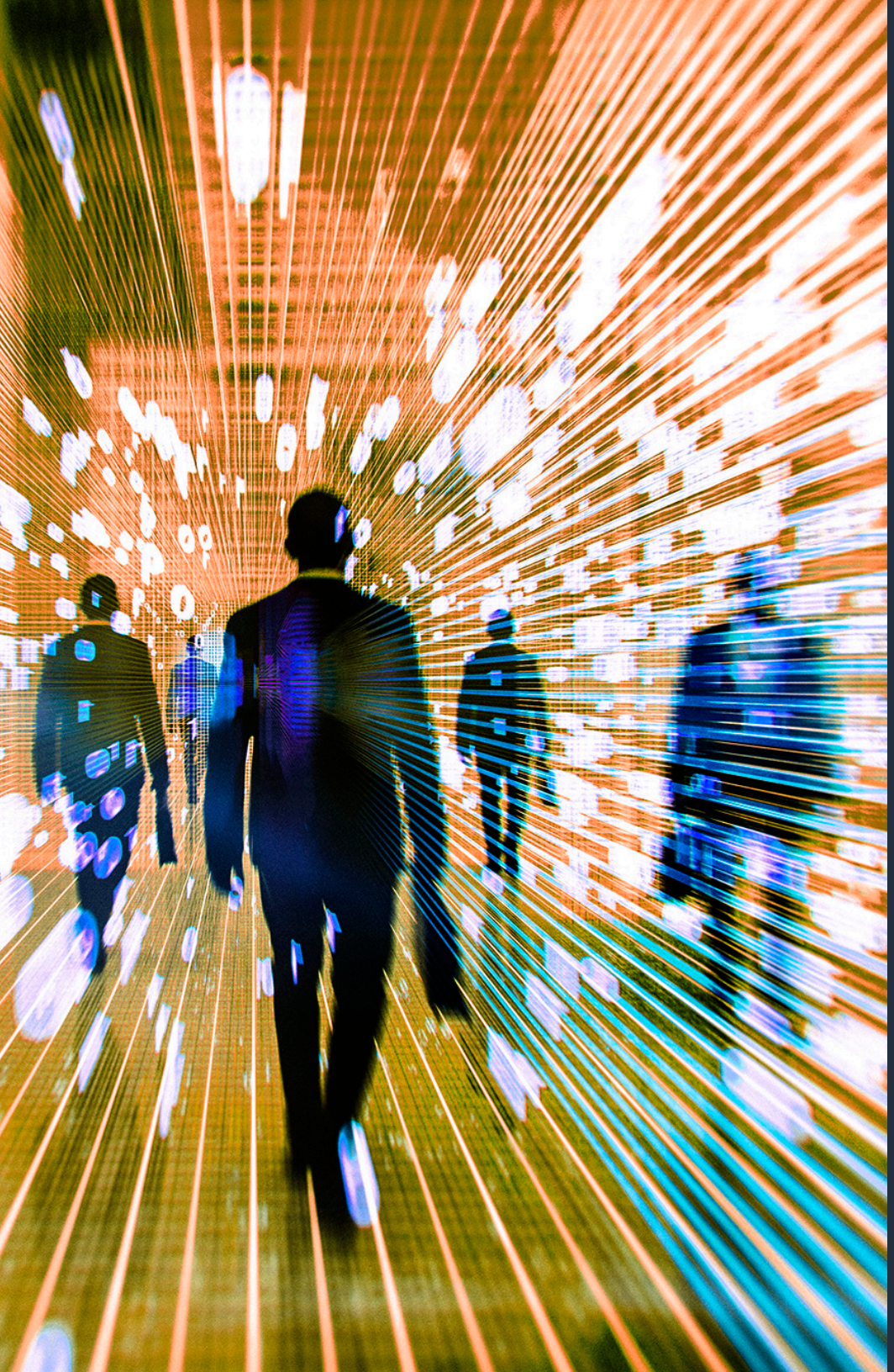
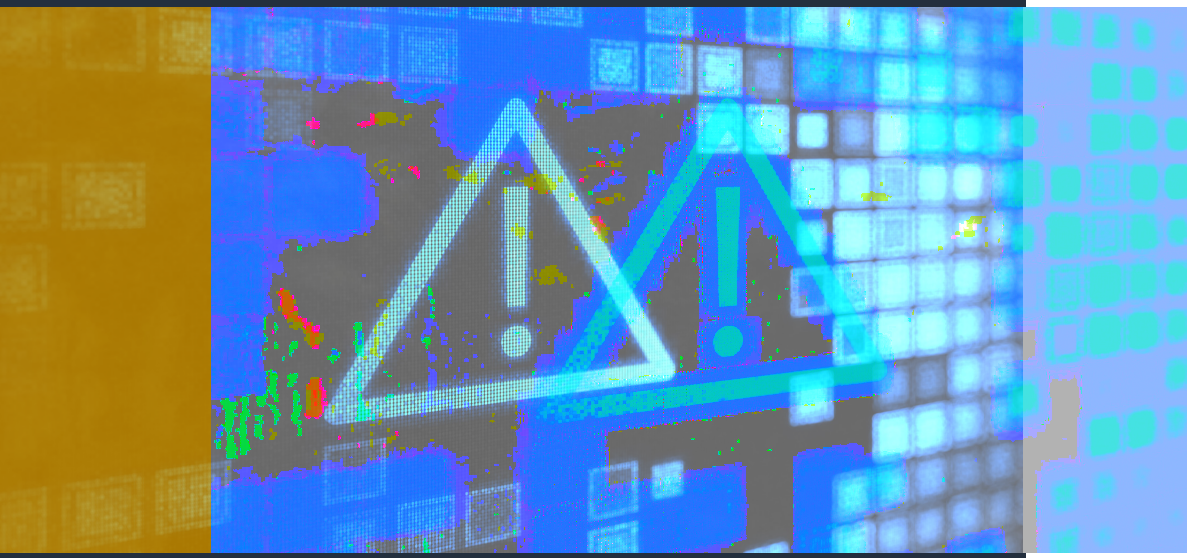


Unmasking your organization's data problem

Joe Chung, Enterprise Strategist and Evangelist at Amazon Web Services

Every company has a data problem



Imagine this...

The weekly Excel report comes out and is delivered to your email. As you review it, you see an anomaly in the financial data that you don't understand, despite the pivot table provided in the report that allows you to drill down to at least some level of detail. You ask your operations analyst what's going on. To which your analyst responds, "I'm not sure. Let me find out."

The next day, the analyst tells you that the reason for the anomaly is productivity was way down at the manufacturing plant.

> "That doesn't make sense," you say. "Can you ask HR if sick days are impacting the productivity numbers? Or could it be that there was an issue with the time capture application at the plant?"
>
> "It will take a week to get at that data and merge it with the financial data," your analyst says.
>
> "Can't you just send me a dump of the data from the ERP and time application, and I'll work with it myself?"
>
> The analyst responds, "I don't have access to the data, and it will take a few days to submit the right tickets to get access to it."

If this scenario seems all too familiar to you, your organization has a big data problem.

Your first reaction may be this is a business intelligence process and tools challenge that has plagued organizations since time began, not a true big data problem. Without getting into a religious debate about what analytics vs. reporting vs. business intelligence are, the key fact is that ***every* organization has a big data problem**. With artificial intelligence and machine learning capabilities starting to come to fruition, it's now more important than ever for enterprises to get a better handle on the data they have, and how to really harness its power to become a **data-driven organization**.

The mental model I have for becoming a data-driven organization is the nervous system of the human body. Nerve endings extend throughout our bodies sending sensory signals to the spinal cord to be processed by our brains and acted upon. It's a model that's being emulated through data architectures that have the ability to receive, process and store data real time from anywhere inside and outside the enterprise. The signals are processed in real time and acted upon by machine learning algorithms. Unfortunately, far too many organizations believe these new capabilities are nice-to-haves, only applicable to specialized data scenarios, or they try to relabel legacy business intelligence platforms as "data lakes."

Data dysfunctions

Most of us think of big data problems as being about volume. The truth is every organization has data issues beyond volume that have been masked in a number of different ways. Here are a few common data dysfunctions that I've observed:

Lonely and discarded data

First, many organizations don't realize that a lot of interesting data is thrown away or is just not accessible. Some examples include data like user activity in the application (and how they might use the application in relation to other applications); the telemetry of the infrastructure hosting the application; or older versions of data that are no longer compatible with current table schemas.

Second, data is siloed across many applications and data warehouses. While no one application may be "big," together, they are big. So, when the business needs to analyze data across multiple sources, it becomes very challenging. That's because siloed data poses an access problem. Every place where data is stored comes with its own access roles, rules, and ceremonies to be adhered to that can make it challenging to access the data.

> Every organization has a big data issue that has been masked in different ways.

Low-fidelity data

Legacy enterprise systems generally process and capture end states and typically only report small snapshots of time. In addition, data is processed in batch rather than in real time. Data can change a lot between batch windows, but older systems are often designed to discard interim changes because they cannot handle the speed in which data can change.

Round data in a square table

Many enterprises realize there are troves of data that don't neatly fit into traditional database storage technologies (for example, images, sensor data, and so on). There is also a lot of variety in how you can analyze and consume insights from the data. For example, when launching a new analytics initiative, you may realize that no single reporting or visualization solution can fulfill all your users' needs. You may need to consider delivering insights processed by algorithms through APIs, within applications using custom visualization widgets using JavaScript frameworks like D3.js, and through business intelligence portals leveraging Tableau and other visualization solutions.

Messy data

Enterprise systems don't like messy data, which is why there are forms, rules, and other validations to make sure data is as clean as possible before it's stored. But some of the most interesting data may not be so clean. When you start tapping into unstructured or object-based data, there's going to be noise. Just like in electronic noise, there are filtering, enhancement, and amplification mechanisms you can use to get at the data you want. One use case that many enterprises are concerned with is the spiraling costs of sending data into proprietary log aggregation, security, or monitoring tools. In most of these cases, there's an opportunity to lower costs by filtering out a vast portion of the log data that isn't useful.

If you've struggled or resonate with any of the points above, it's time your organization revisits your analytics approach and architecture. Every enterprise has the opportunity to deploy fit-for-purpose analytics solutions (storage, processing, querying, analysis, presentation, and so on) to meet their existing business and IT challenges.

Modern analytics platforms enable critical business insight

Once you're ready to tackle your big data problem, what can you reasonably expect to accomplish with a modern analytics platform? Here's what's possible and how it's being done today, from a technical perspective.

> "
>
> Data-driven decisions necessitate access to many disparate types of information.

Access to any data I want

Data-driven decisions necessitate access to many disparate types of information. A pilot relies on the gauges in the airplane to understand information critical to flight—like altitude, air speed, fuel consumption. But imagine if the pilot didn't have those gauges all in one place. Perhaps they have to walk to the back cabin or radio in for the information or worse yet, have to ask permission for the data. Unfortunately, this is a daily reality in today's enterprise environment.

Forward-leaning organizations have flipped this standard on its head by pulling data out of the systems in which it exists and storing it in one place (that is, a data lake). While there are many instances of companies storing large amounts of one type of data, more and more companies are creating enterprise-wide data lakes that contain multiple data types from different sources.

Internet-scale companies like Amazon, Yahoo, and Facebook started to see in the early 2000s that relational database technologies had reached their limits in terms of scalability and performance. Amazon responded with a technology called Dynamo, which is a highly available and scalable key value store, such as NoSQL/non-relational technology. Amazon then evolved and leveraged Dynamo to create services such as Amazon S3 and Amazon DynamoDB. Amazon S3 is attractive to enterprises looking to create data lakes due to its ability to store many different data types and its low storage cost. There are, of course, other technical solutions, including Hadoop, but an important characteristic of all data lake solutions is their ability to store all types of data at petabyte scale and at low cost.

Responsive to change

Business systems and data change all the time, but often the systems which report or share that information end up being the last to change. How many times have you been told it will take six months or more for data to be remediated in data warehouses and reports? Or that data changes from source systems have not yet flowed to reporting systems, and that it takes several days for those changes to make their way down due to the batch processing? **The speed at which data is available dictates the speed at which decisions can be made**. Therefore, we should expect that modern analytics systems be able to process and report data in near real time and be responsive to changes to upstream data sources.

"

The speed at which data is available dictates the speed at which decisions can be made.

The first key enabler is the nature of how data is stored in big data technologies like Amazon S3 or Hadoop. One of the big inhibitors to changing a relational database is modifying the schema or definition of how data should be stored. Until the schema is modified, data can't land in the database or it will break. File or object-based technologies like Amazon S3 don't care how data is structured—data can come as it is versus the "you need to fit my structure" approach.

The other challenge is that only one schema is active at any given time. While I'm sure we've all seen database tables named "2015" and "2016," it's not ideal. Big data technologies have a schema on read-based approach, meaning the structure of the data is applied when you grab it and not inferred based on how it's stored. What that means for businesses is that data changes from source systems aren't a big deal.

The second enabler is streaming technologies like Amazon Kinesis and Apache Spark. Most enterprises move data around in big batches; typically, this occurs once a day. Streaming technologies allow data to be ingested in smaller pieces at a massive scale. For example, SONOS, the speaker manufacturer, processes 1 billion events per week using Amazon Kinesis. You should never have to wait for the daily batch to be completed to understand where your business stands.

Interactive insights where and how I want them

Today's enterprise users jump through lots of hoops to understand the information being presented to them. Maybe it's digging through their inbox to find the report that was attached. Or logging into the reporting system to download a PDF only to find they have to copy and paste the data into Excel to make sense of what is there. We need to stop making users slog through terrible experiences to get the data and insights they need. The rallying cry for users should be: Bring the data in the right form with the right tools at the right time.

Software like Tableau, Amazon QuickSight, and others have made things better by considering user experience when interacting with data. However, I have found that at most enterprises it requires the use of many tools to meet users' needs. It could be Amazon QuickSight embedded in a business intelligence portal to a Tableau workbook sent in an email. AWS provides the diversity of data storage and business intelligence tools delivered through a pay-as-you-go model. This allows organizations to experiment with many different business intelligence tools without making large investments in infrastructure and licensing.

We need to stop making users slog through terrible experiences to get the data and insights they need.

aws

> The best algorithm in the world is worthless unless it can be integrated with business processes.

One persona you should not ignore is the data scientists in your organization. Jupyter notebooks have really taken off in the data science community; they are part content management, part code execution, and part visualization. It's a very powerful tool to share knowledge and document and execute machine learning algorithms. Amazon SageMaker is a managed notebook environment that takes care of all the heavy lifting for you and your data scientists

Intelligence embedded in the business

Artificial intelligence and machine learning are all the rage these days, and rightly so. Advances in machine learning frameworks coupled with the use of specialized servers utilizing graphics processing units (GPUs) are enabling all kinds of new capabilities, like autonomous driving. Of course, in order to train machine learning models vast amounts of data are required (thus the data lake points I discussed above). Organizations are already starting to take advantage of these AI/ML capabilities to drive new outcomes not previously possible, such as being able to better predict health outcomes based on retinal imaging, or predict outages or hardware breakdowns in the field. Businesses can strengthen their organizational AI/ML muscles by letting AWS take care of the heavy lifting, as this is not the stuff of science fiction but operating in production today.

One final point I would like to emphasize is that the best algorithm in the world is worthless unless it can be integrated with business processes. Often getting the insight or data science model created is the easy part—getting it integrated into your insurance policy engine or retail platform is the hard work, as these systems do not typically have the ability to integrate outside data sources or APIs. This is a great opportunity to consider moving these systems to the cloud to take advantage of all the services available to help modernize or re-architect them.

Organizing for insights

Building an advanced analytics capability at your company is not just about the technology. Often, the biggest challenge organizations face has to do with the organization—its processes, governance and people. So, what do you need to keep in mind as you organize your analytics investment for success?

Start with an analytics Center of Excellence

To go from strategy and intent to meaningful progress, one of the first steps is to identify and choose a leader and a team to spearhead the change, and to set up an analytics Center of Excellence (COE). The team typically starts small, with a few cross-functional roles to bootstrap it; the team will then grow to service more and more needs.

Many large enterprises already have established shared services organizations that handle business intelligence or reporting. Those organizations can seed the analytics COE with technical and business roles. Similar to IT infrastructure organizations, they should not only supply the talent, but also be a key driver and sponsor for the effort. Because over time, these reporting shared services organizations will need to evolve to either adapt or become part of the analytics COE. The starter roles are often data engineers and architects, business intelligence analysts, and data scientists. The group needs to be led by someone who can work across multiple organizations, business units, and back-office groups, such as Finance and IT.

Serve all your customers' needs

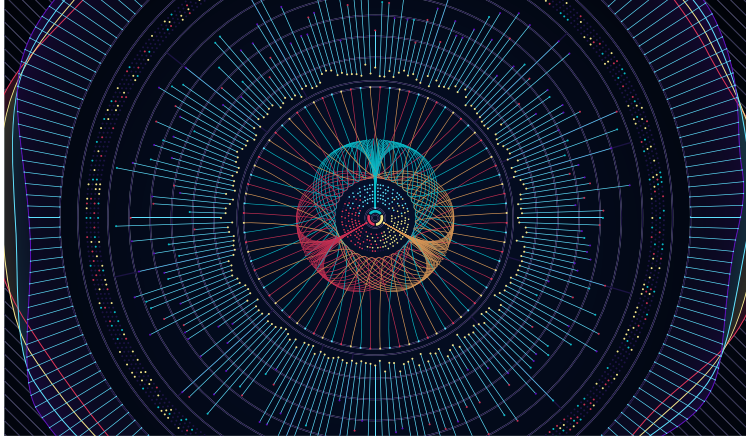
One of the primary mental shifts that organizations need to make is moving from an attitude of "you must use our reporting solution and you will like it" to "what are your analytics needs and how can we help enable you?" Shared services reporting organizations are often report pushers and are not positioned to answer the difficult questions that come from employees, business leads, and customers.

Therefore, when establishing a new analytics COE, it's important to establish tenets for the group which will set expectations for how the group acts and makes decisions.

The analytics COE will need to service two types of customers:

- **The data and analytics consumers**: The decision makers, data scientists, business intelligence (BI) analysts, and developers. These customers typically care about the ability to quickly access insights and data and the quality of the tools and services available to them to process and present data.
- **The data producers**: The owners of applications, infrastructure, and devices who will supply the data into the platform. These customers need services like the ability to easily publish their data into the analytics platform and define a data contract. This includes the domain model of the data, frequency of refresh, and definition of policies, for example, a security policy outlining who can access their data.

It's not just about having the latest tools in the belt; it's about making it easy for your customers to get what they need.



The analytics capability and the platform need to service both types of customers—if their needs are not met then the analytics effort will not deliver the business value. Therefore, it is critical to have a mechanism to capture the needs of these two types of customers across a potentially very large and diverse set of business units and personas. Some organizations set up advisory boards or work with a few key stakeholders to drive the need. There is no single correct answer, but having a mechanism to capture the voice of the customer(s) and prioritizing their needs is critical.

Rethink the COE

An analytics COE services and presents a specialized set of cloud services focused on meeting analytics needs. In the past, reporting and BI organizations often provided one solution to fit everyone's needs (a one-size-fits-all strategy). In this era of rapidly evolving technologies in big data, rich visualizations, automated decision making, artificial intelligence, and machine learning it's just not possible to have a single technology stack. It's not just about having the latest tools in the belt; it's about making it easy for your customers (producers or consumers) to get what they need.

COEs run the risk of becoming concierge services. That can be fine for certain types of requests, but the COE can quickly get overwhelmed and backlogged with requests if there's no scalable, self-service mechanisms and transparent prioritization and governance processes in place. Analytics COEs need to engineer and architect a self-service, secure, operable, and scalable data platform with an ever-evolving ecosystem of technologies to process, analyze, and present insights.

While becoming a data-driven organization won't happen overnight, pinpointing your data challenges, organizing a plan for how you will serve your customers, and empowering your teams to deliver the right value at the right time are steps in the right direction.

About the Author

Joe Chung is an Enterprise Strategist and Evangelist with Amazon Web Services.