NAVIGATING THE SEAS OF DATA:

AWS Data Warehousing Strategies for Smooth Sailing and Deep Insights with CDW

CONTENTS:

What we mean by data 4

Awash in a sea of data 6

Setting sail on the lake 8

To sail or to warehouse 11

Lifeblood of Business Intelligence 13

Data warehouse architecture 14

Accelerated insights with Amazon Redshift .. 16

Conclusion 16

“Water, water everywhere, nor any drop to drink.”

So said the titular character in *The Rime of The Ancient Mariner*, a poem by Samual Taylor Coleridge. We use this phrase to describe situations where we find ourselves surrounded by something we cannot benefit from. This seems like a fitting description for many organizations to describe their relationships with their data.

WHAT WE MEAN BY DATA

In the broadest sense, we use the word ***data*** to describe a related collection of values, discrete or continuous, that can convey information. For data to convey information, someone must evaluate it for meaning. We can use data to convey meaningful information about such things as quantity (e.g. gallons of fuel in a tank), quality (e.g. customer satisfaction), statistics (e.g. frequency of replacing a machine part), and facts (e.g. zip codes). We use data to reflect and evaluate nearly every aspect of our lives.

In computing, data simply describes a sequence of related symbols. Computers operate in binary, using ones (1) and zeros (0) to represent symbols digitally. Thus, in the context of computing, all data gets stored and manipulated as a series of digital electrical or optical bits. Since we live in an analog world, computers must convert analog signals into digital signals first using an analog-to-digital converter.

Consider the digital camera in our phone. Analog optical information (photons) of varying intensities and wavelengths from outside our phone enter the camera lens and interact with an imaging sensor composed of a grid of photosensitive microscopic dots (pixels). Each pixel generates electrical voltages when exposed to light. The sensor assigns digital numerical values to those voltages for each pixel location on the grid symbolizing intensity and wavelength. This may look something like this:

Location: 1133, 248
Red: 13
Green: 119
Blue: 243

The digital image data acts as a map of all the pixels and their values. For each pixel of the image, the digital image data includes key, context-relevant information: its X,Y location on the grid and the intensity levels on a scale of 0-255 for each Red, Green, and Blue (RGB) wavelength. The collection of all the key point-in-time values from all the pixels working in tandem on the imaging sensor grid make up the data of our digital image. The digital image not only contains the data captured but also **metadata**. Metadata refers to data about data. For a digital image, metadata may include the location where we captured the image of an event, the date and time of the capture, the equipment used to capture the image (camera, lens, F-stop, etc), and other pieces of relevant information.

We can use data to convey meaningful information about such things as quantity, quality, statistics, and facts.

Organizations generate more data than they can capture and store and they capture and store more data than they can use.

AWASH IN A SEA OF DATA

Organizations of all shapes and sizes generate and use vast amounts of data in the normal course of their work. Data can come from a variety of sources both outside and inside the organization. Outside sources could include supplier lead times or weather forecasts. It can come from inside sources such as sales transactions and employee punch-ins/outs. Generally speaking, organizations generate more data than they can capture and store. They also capture and store more data than they can use.

Data comes in many different types: analog or digital, continuous stream or discrete events, structured or unstructured, realtime or historical, etc. What works to store and move one type of data may not work for another type. Solutions to store and move disparate kinds of data can prove complex and costly to implement and operate. Before we embark on engineering a solution to capture, store, move, and process a particular kind of data, we first need to examine two fundamental areas: relevance and feasibility.

Relevance

As previously mentioned, data requires people making meaningful evaluations to act as ***information***. We use information for decision making. How might this particular data inform our decision making? If we have no clear answer, it may not make sense to invest in a solution to use it. Simply because we can do something doesn't imply we should. Often, however, we have already identified an information gap in our decision-making process that a particular type of data would fill, making the answer to that question seem fairly obvious.

Feasibility

After the question of relevance comes feasibility. Feasibility has several dimensions to it:

- **Capture**
  Do we have the ability to capture the data we need? If we want to capture analog data, what kind of analog-to-digital conversion do we need to do? Can we use a converter already available, or will we need to custom build one? For capturing purely digital data, do we need any software connectors or custom interfaces to connect disparate data sources and systems?

- **Storage**
  Do we have the capacity to store and retrieve the data we capture? Can we store all the data we capture or just a subset? Do we need to retain the data for historical or compliance purposes? Can we retain it long enough? Can we store the data as quickly as it's generated? If not, how can we solve for lag time? Can we scale the storage to keep up with data growth? Can we store the data securely? Can we protect the data from loss?

- **Transport**
  Do we have the ability to move the data from one location to another? Do we need redundant transport pathways? Can we solve for delays or bottlenecks in data movement that may impact our operations?

- **Process**
  Do we have enough computing power to process the data for meaningful insights? Can we process the data in a timely enough manner to act on the information it provides?

Feasibility means not just evaluating technology but also examining monetary considerations. We must justify the cost of our initial investment required to capture, store, transport, and process the data we seek. We also must justify the ongoing costs to operate our solution (licensing subscriptions, support contracts, staff salaries, etc). Compare these to the intended operational improvements the insights of this data will provide us in terms of cost-savings or revenue-growth.

SETTING SAIL ON THE LAKE

We've seen how data comes in a variety of types. Two primary differences concern ***structured*** as opposed to ***unstructured*** data. Structured describes highly organized data that follows a predefined format or schema. Consider databases and spreadsheets as examples of structured data. Unstructured data lacks a predefined schema and can include things like images, videos, emails, freeform text documents, and other types of data that don't fit well into a predefined structure like the rows and columns of databases and spreadsheets.

Organizations often want to store large amounts of raw (unprocessed) unstructured data or a mix of unstructured and structured. For this, we turn to a ***data lake***. Data lakes are designed to ingest and store data in its native form, allowing for flexible exploration and analysis. Data lakes have several characteristics to consider.

Ingestion

No, we're not talking about a sour tummy after eating a big meal. Ingestion refers to how we get the raw data into the data lake from a variety of sources: IOT sensors, surveillance cameras, social media feeds, etc. Here, one of the primary areas of focus concerns whether to continuously ingest data in real-time or in periodic batches like the end of a business day or financial reporting period.

Storage

Scalable data storage forms the foundation of any data lake solution. Organizations have many different considerations to evaluate for their storage layer, including:

- Public cloud, on-premise, or hybrid
- File, block, and/or object
- Hardware-based vs software-defined
- Performance vs capacity optimization

Not only do we need scalable storage, we often need it distributed across multiple systems or locations. A distributed file system can help satisfy this need, such as the popular Hadoop Distributed File System.

Integration

A data lake integrates data from multiple sources, combining it into a single repository. This facilitates a consolidated view of data from disparate sources. Storing raw (unprocessed) data reflects one of the primary purposes of a data lake, but occasionally data must undergo an **Extraction-Transformation-Load** (ETL) process to read data out of a data source for ingestion into a data lake.

Not only do we need scalable storage, we often need it distributed across multiple systems or locations.

Schema-on-read

We have discussed that data lakes store large amounts of unstructured as well as structured data. We want to build our data lake solutions to store data as near to its original raw, unprocessed state as possible. When we ingest the data into the data lake, writing to the data storage, we do not want to enforce any particular data schema or structure. Upon reading the data from the lake to perform analytics and interpretation, users apply a schema based on their particular needs. This enables diverse users to perform various kinds of analysis without having limitations enforced during the data capture which could impede.

Governance

While not a building block of a data lake solution per se, effective data governance policies play a crucial role for organizations to get the most benefit from their data lake. We often see organizations building a data lake to capture and store any and all data possible even before deciding if and how they might use it. They reason: better to have it and not need it than to need it and not have it. This can spiral out of control, however, and the data lake becomes a so-called "data swamp". The data governance conversation should not begin after implementing a data lake solution but rather in the planning and design stage.

Governance not only addresses what kinds of data to capture and store but also how to secure it. Different types of data have different needs for security. Thermal sensor data from a warehouse, for example, has very different requirements for security than personnel data like employee punch-ins and punch-outs. Also, data in transit (i.e. moving across a network connection) has different security considerations than data at rest (i.e. stored in the data lake).

TO SAIL OR TO WAREHOUSE

We often encounter people using the terms data lake and ***data warehouse*** interchangeably. While they share many of the same characteristics, they differentiate in one key way: **unstructured vs structured data**. For organizations looking to store, analyze, and inform their decision making using large amounts of structured data, a data warehouse solution fits the bill.

For review: **structured data** describes data formatted and organized in a specific way (a schema), which makes it easy to read by both humans and machines. The data that we ingest into a data warehouse can come from a variety of structured data sources such as databases, spreadsheets, and specially-formatted text files (e.g. JSON, XML, YAML, etc). Most organizations' business systems generate and use structured data. Consider the booking system of a hotel

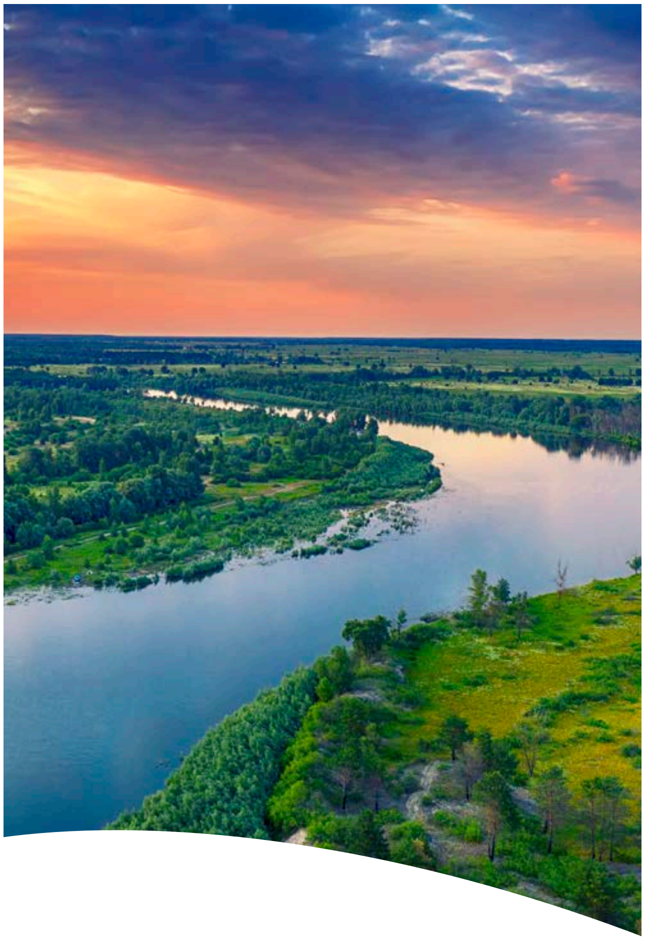
> ***For organizations looking to store, analyze, and inform their decision making using large amounts of structured data, a data warehouse solution fits the bill.***

chain or the point-of-sale system for a retail store. Structured data doesn't just come from transactional systems, either. A hospital's electronic health records, for example, don't reflect transactional operations the way a retailer's point-of-sale records do.

With structured data, we assign each data element a particular field or column. Each record, or row, represents a specific instance of that data. An organization may have records for all their customers that include fields for the customer name, billing and shipping address, phone number, email address, and other uniquely identifiable information. They may also have inventory records that include a product number, a description, a purchasing cost, a selling price, and how many items they have on hand. When a customer places an order, the point-of-sale system queries the customer database for shipping and billing information, queries the inventory database for product information, queries the financial systems for customer credit terms or outstanding balances, etc. After the sale, their point-of-sale system then updates the inventory system to reflect an accurate quantity-on-hand for the items purchased.

Organizations treat their structured data as a highly valuable resource due to the ease of searching, querying, and analyzing it for business insights. It feeds data-driven applications such as Business Intelligence (BI), data mining, and advanced analytics. Artificial intelligence and machine learning (AI/ML) models also benefit from structured data. The schema-on-read approach for data lakes does not apply to data warehouses. The extraction operation to read the data from the source and ingest into the data warehouse includes extracting the fields or data descriptors (keys) from the data source.

Sometimes the pre-existing data schemes in our data sources do not adequately provide for the kind of analytics we need from our BI or other analytics solutions. In these situations we employ a similar ETL process we spoke about previously in the context of data lakes. We **extract** (read) the data and schema from our source. Next, we process that data to **transform** it to a usable format prior to ingestion into the data warehouse. This could mean something as simple as changing data descriptors or keys or making a series of complex calculations on the data and storing those calculations. Finally, we **load** (write) the newly-transformed data into our data warehouse solution. Additionally, a data warehouse may not only include data extracted from a data source but also its metadata to help enable further insights that the data alone could not. Metadata expands the context of the data captured and analyzed.



LIFEBLOOD OF BUSINESS INTELLIGENCE

Organizations implement data warehouse solutions primarily to enable and support BI initiatives. BI refers to strategies and tools that organizations use to analyze and manage their business information*. BI encompasses activities such as operational reporting, data mining, predictive and prescriptive analytics, performance management and benchmarking.

Organizations use BI to help inform a wide range of business decisions, both operational and strategic. BI tools empower organizations to gain insight into new market opportunities, to assess product demand or quality issues, to gauge the impact of sales and marketing efforts, and a myriad of other uses. From a small nonprofit to the largest global enterprise, everyone can benefit from more informed business decisions.

Business Intelligence tools empower organizations to gain insight into new market opportunities.

*Recall the difference between data and information

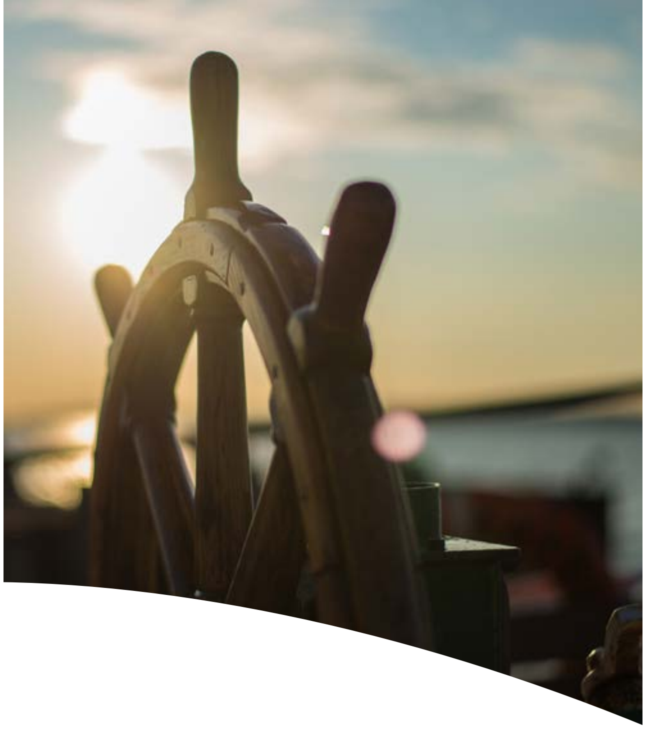
DATA WAREHOUSE ARCHITECTURE

Organizations wishing to embark on a journey to building and operating their own data warehouse solution have a few things to consider, starting with the architecture of the solution. Data sources, ETL processes, storage technologies, and data/metadata management all factor into what kind of solution architecture would best serve an organization's needs. Additionally, organizations must consider budget, existing infrastructure, and any future scalability requirements.

Data warehouse solutions reflect three primary architectural approaches: the traditionally centralized **Enterprise Data Warehouse** (EDW), a decentralized **Data Mart**, and a modern **cloud-based data warehouse**. Each approach has distinct advantages and disadvantages.

Enterprise Data Warehouse (EDW)

The traditional EDW has a centralized architecture, utilizing a common data repository for data. Data gets extracted from various structured data sources and transformed through direct processing before getting loaded into the repository, usually a relational database (RDBMS). Advantages to this approach include having a centralized view of the data to ensure consistency, support for long-term data retention for historical analysis, and robust access controls and security. Disadvantages include large upfront costs, scalability challenges, and the potential for added latency from ETL processes to negatively impact performance.



Cloud-based solutions support real-time ingestion and analytics to handle streaming data sources.

Data Mart

A Data Mart solution replaces the centralized approach of an EDW solution with a decentralized approach. A data mart contains a smaller data set specific to a particular subject. Organizations build data marts to serve the needs of specific functional areas or business units. ETL processes extract data from disparate data sources or from a centralized repository like an EDW, transform the data for subject-matter specificity, and load the data into a smaller data mart repository. The primary advantage of this approach compared to an EDW reflects smaller startup costs and faster implementation. Data marts bring easier management and ongoing maintenance due to their smaller size. They align to the specific needs of specific business groups. However, the smaller size and increased number of data marts can introduce challenges with data consistency across different marts: a primary disadvantage compared to EDW solutions. This can lead to duplication of efforts and wasted resources from redundant data as well as difficulties integrating data from different marts for an enterprise-wide analytics.

Cloud-based Data Warehouse

In the cloud computing era, more and more organizations have turned to public Cloud Services Providers (CSP) for data warehousing solutions. These cloud-based solutions leverage the infrastructure and services in the CSP's catalog to store and process data. CSPs build these solutions using distributed and scalable storage systems. They often support a mix of both unstructured and structured data, blurring the traditional boundaries of a data lake and a data warehouse. Cloud-based solutions offer several advantages over on-premise solutions. For starters, the pay-as-you-go consumption pricing makes them very cost effective, eliminating the steep expenses associated with upfront hardware and software costs. Their elasticity and scalability provides organizations with the flexibility they need to adjust to changing demands. Cloud-based solutions support real-time ingestion and analytics to handle streaming data sources and easy integration to other CSP services and tools. Some organizations may consider the security implications of cloud-based storage as a disadvantage compared to on-premise solutions. Furthermore, cloud-based solutions may introduce additional latency or other performance bottlenecks. They also may require redesigning any ETL processes. Finally, some organizations perceive any potential difficulties in switching to a different CSP vendor as disadvantageous vendor lock-in.

ACCELERATED INSIGHTS WITH AMAZON REDSHIFT

As the market leader in cloud computing, Amazon Web Services (AWS) provides organizations with the services they need to achieve their objectives quickly, reliably, and cost effectively. Organizations looking for a cloud-based full-featured data warehouse solution will find what they need from AWS with **Amazon Redshift**.

With Amazon Redshift, organizations do not have to commit large capital expenditures to build out a data warehouse. They do not need to invest in data warehouse management expertise and skills. They can rapidly go from data to insights, enabling them to deliver on their desired business outcomes without worrying about provisioning and managing their data warehouse.

AWS has taken any guesswork out of starting the data warehouse journey. Amazon Redshift Serverless automatically provisions resources, intelligently scaling data warehouse capacity. This provides tremendous flexibility considering that data warehouse workloads can have unpredictable performance demands. Before cloud-based data warehouses, organizations had to make an educated guess about how much storage capacity or processing power they would need to build into their solution and whether or not to build for peak demand. Further, organizations do not incur charges when the data warehouse sits idle, only paying for what actually gets used. With Amazon Redshift Serverless, data analysts, developers, and data scientists can get insights from data in seconds by loading data into and querying records from the data warehouse.

Organizations that have higher capacity or processing demands than Amazon Redshift Serverless provides can manually provision their own Amazon Redshift data warehouse. An Amazon Redshift data warehouse is a collection of cloud computing resources called nodes which get organized into a group called a cluster. Each cluster runs an Amazon Redshift database engine and contains one or more databases. AWS offers a variety of compute nodes to accommodate different workloads. Organizations who chose this option will require additional expertise and skills in managing a data warehouse but will benefit from pay-as-you-go billing, quick and easy resizing, and the global scale and reach of the world's leading CSP.

CONCLUSION

Data literally surrounds and engulfs every organization across every industry. To turn that data into ***information*** which informs and enables improved decision making, organizations turn to data warehouse solutions for their structured data analytics. Cloud-based data warehouse solutions like Amazon Redshift from AWS can help those organizations accelerate the time it takes to realize value from the insights their data can provide in the most efficient, flexible, and cost-effective way.

To find out more about how your busines can gain value insights from your data with Amazon RedShift, visit:

CDW's AWS Redshift Migration Accelerator ›