Zetaris
The Networked Data Platform

CRUSH YOUR **CLOUD COSTS**

JUNE 2021

ENTERPRISES HAVE MOVED AND ARE MOVING THEIR DATA TO THE CLOUD FOR COMPELLING REASONS.

But data egress fees and the cost of analytics workloads can significantly erode the cost savings expected from cloud migration. Through generic illustration, based on published research, this paper shows how adoption of Zetaris: The Networked Data Platform can reduce the rate of this erosion, ensuring that the TCO benefits of cloud migration can be realized.

Enterprises have moved and are moving their data to the cloud for compelling reasons. It reduces expenditure on IT hardware and facilities. It provides flexible (i.e. consumption-based) payment options. It promotes enterprise-wide mobility and collaboration. Back-up and disaster recovery are outsourced to the cloud provider.
However, enterprises can receive a nasty shock when they receive their cloud provider's first invoice. This is for two reasons, data egress and analytical workloads. Zetaris explicitly addresses this with an Intelligent Adaptive Cache.

1. The Intelligent Adaptive Cache

A unique feature of Zetaris: The Networked Data Platform is the Intelligent Adaptive Cache. The Intelligent Adaptive Cache analyses your cloud workload and automatically implements optimisations, including intelligent local caching of commonly accessed data. Frequently used queries are cached to improve data retrieval performance by reducing the need to access the slower underlying storage layer.
Star schemas, a data modelling approach commonly used in cloud data warehouses, consist of fact tables and dimension tables. The Intelligent Adaptive Cache of Zetaris: The Networked Data Platform caches dimension tables into main memory to remove duplication and improve JOIN performance. It also caches fact tables to improve query performance. The Intelligent Adaptive Cache determines dimension and fact tables automatically from the query pattern to reduce query time and implicitly cache them.

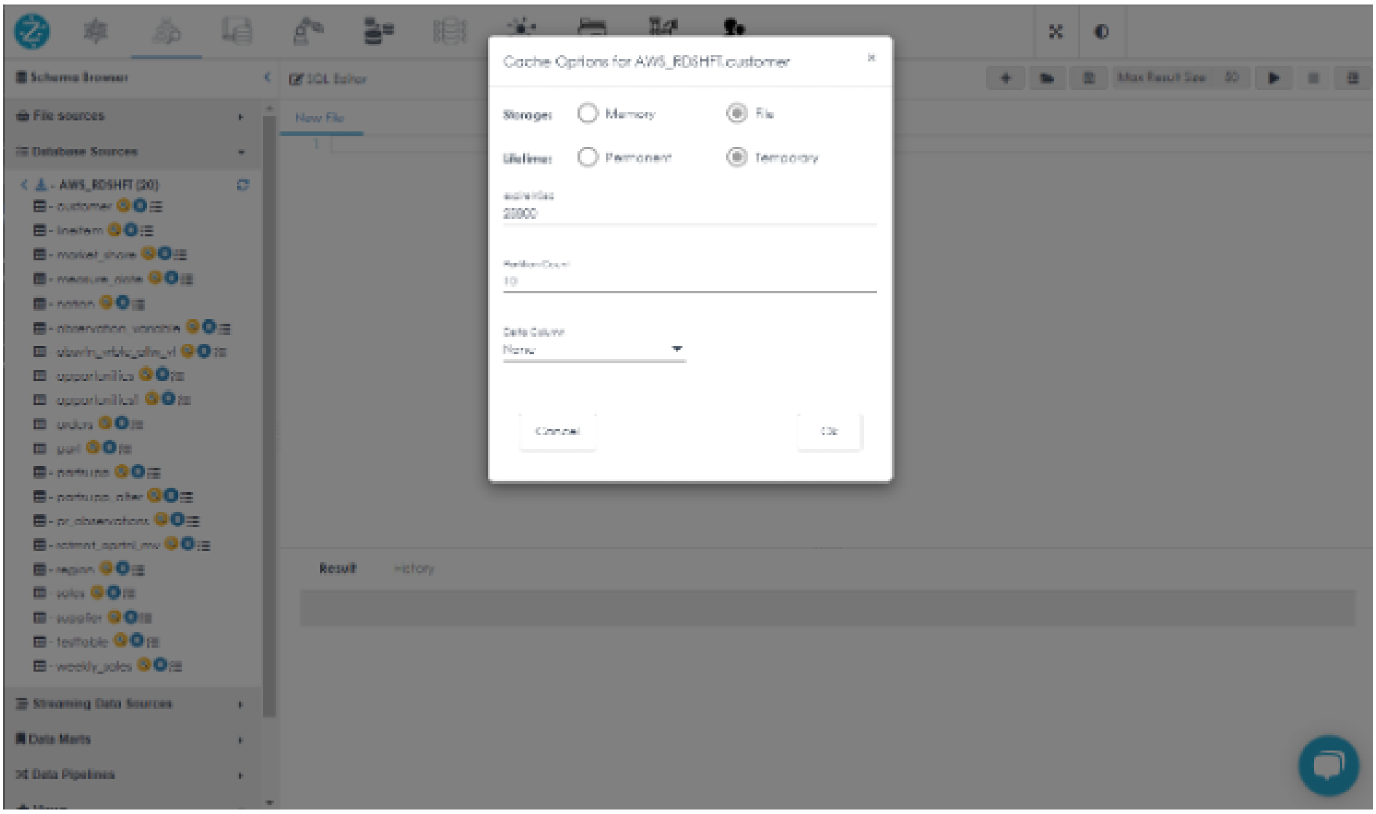


Figure 1: Zetaris - The Networked Data Platform Screenshot showing the Intelligent Adaptive Cache Options Dialog

1. Data Egress

Data egress fees are one of the cloud's biggest hidden costs.

In simple terms, data egress refers to data moving out of the cloud environment. Most cloud providers allow you to input your data (ingress) at no cost but will charge large network fees to move your data out of their environment (egress).

It is common for cloud providers to charge between 5 and 20 cents per GB every time you move your data from their cloud storage to your on-premises location or to another cloud environment. Major businesses pushing and pulling terabytes of data to and from the cloud can expect very large monthly invoices.

For example, if you are transferring 1 TB of data from your cloud provider's environment to somewhere else, you could be looking at a cost of $100 per transfer. That may not sound like much but, if you are performing the transfer ten times per day, your monthly bill will be in excess of $20,000. (If your data is moving overseas the cost will be much higher.)

Data transfer fees are determined by a number of factors. These include:

- The amount of data transferred,
- Whether the data is transferred within a region or across regions,
- Whether the data is transferred within or outside the availability zone, and
- Whether the data is transferred within or without the provider's cloud.

These variables, combined with the unknown timing of data movements, make it difficult both to estimate and to forecast data transfer costs with accuracy.

2. Analytics Workloads

Analytics workloads are sets of queries that read and write datasets. They are fundamental to competitive advantage because they yield results that tell the enterprise, for example, what factors influence customer retention and what factors are linked to customer churn.
However, analytics workloads are very hungry in terms of compute, storage and memory resources, all of which will be line items on your cloud provider's monthly invoice. They can also drive up bandwidth costs if you require faster response times. (Bandwidth is another line item on the monthly invoice.) This is because the cloud takes a brute force approach to solving performance problems: it adds elastic resources on demand. The approach undoubtedly improves query performance but it escalates cost—often dramatically.

To reduce this cost, IT specialists within the enterprise will endeavour to optimize the queries. Analytics workloads typically involve complex queries; for example, 'find all employed male customers aged between 25 and 40 who earn less than $X and who purchased product Y between 1 January and 31 March this year'. The structure of the query plan, meaning the sequence of steps required to retrieve the necessary data, will influence directly how many input-output operations will be performed, the CPU path length (i.e. the number of machine instructions needed to execute the query), the amounts of buffer space and storage time required, and the types of data join required. Query optimization, in essence, is about selecting the most efficient plan for a particular query. Effective query optimization requires highly sophisticated algorithms and a skilled and vigilant database administrator. It is labour-intensive and, importantly, database-specific.

3. A Generic Illustration
What follows is a generic illustration of the effects of data egress and analytics workloads on cloud costs.

Z-Corp is a large enterprise that migrated some of its data to Cloud Provider A and some of its data to Cloud Provider B while retaining its most sensitive data on-premise.

The benefits to Z-Corp were:
- A reduction of hardware and facilities costs,
- A reduction of IT staff costs,
- Productivity gains through the adoption of SaaS models, and
- A reduction of software licensing costs.

Dollar figures are notoriously difficult to establish in this area because enterprises do not make specific figures available. However, based on published research, we can say that the movement to the cloud was calculated to reduce Z-Corp's TCO by approximately 30%.

The literature presents cloud benefits in qualitative terms: the quantitative benefits are asserted rather than supported with financial evidence. There are studies which compare the pricing of various cloud providers, but these are of little value because they do not provide real-world examples of how these pricings relate to the IT architecture and data-derived activities of specific enterprises.

Of course, the effort involved in moving large volumes of data to the cloud is very expensive. If we assume that Z-Corp's TCO was $15M before the shift to the cloud, a standard ROI projection for the exercise (round figures for simplicity) would look something like this.

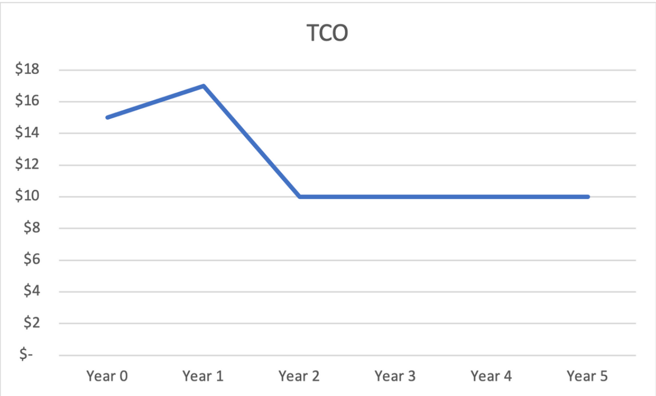


Figure 1: Z-Corp's TCO after movement to the cloud (projected)

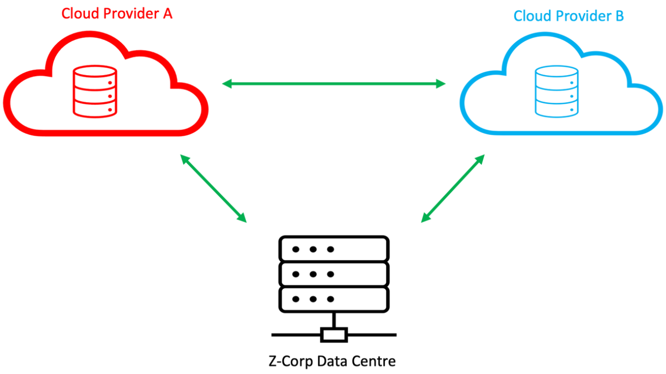


Figure 2: Z-Corp's data landscape after movement to the cloud

From Year 2 onwards, however, Z-Corp's projected cost savings were increasingly eroded.
Moving data into the environment of Cloud Provider A did not cost Z-Corp anything but moving it out of the environment of Cloud Provider A to that of Cloud Provider B and to the on-premise Data Centre costed a great deal. The same applied to data movement from Cloud Provider B. Similarly, despite the best efforts of Z-Corp's database administrator, the analytics workloads proved to be very expensive.

Over subsequent years, Z-Corp's cloud costs rose as follows.

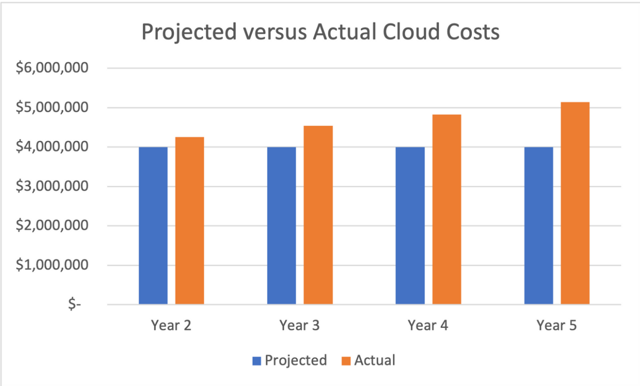


Figure 3: Z-Corp's projected and actual cloud costs

Cost increase has been averaged out at +6.5% per annum based on industry studies.

In other words, Z-Corp's projected cloud spend was over $1M higher than estimated by Year 5. The effect of this annual increase on Z-Corp's TCO is indicated below.

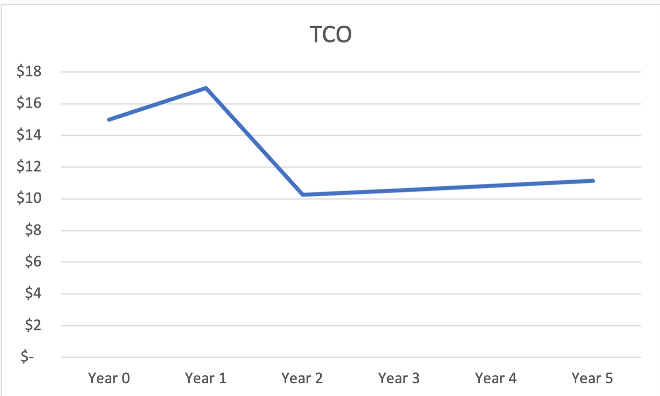


Figure 4: Z-Corp's TCO after movement to the cloud (actual)

4. The Zetaris Solution

Zetaris: The Networked Data Platform addresses specifically the problems of data egress and analytics workloads in the cloud. Its speed and scalability derive from virtualization over metadata, which allows you views of any combination of data from disparate sources without having to move the data itself.

4.1 Reducing data egress

Zetaris: The Networked Data Platform ingests the metadata from disparate and disparately located data sources to construct a virtual schema. The virtual schema configured allows these disparate data sources to be interrogated as if they were physically consolidated in a data warehouse. The data itself, however, remains where it is.

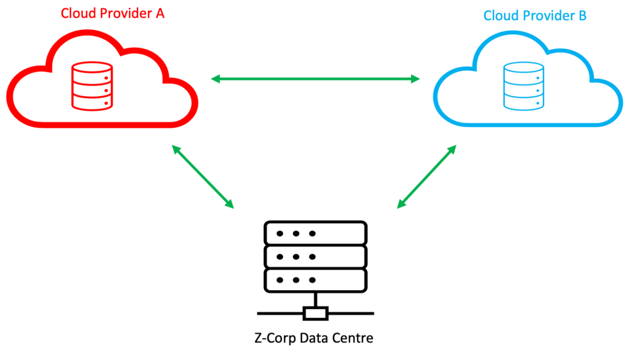
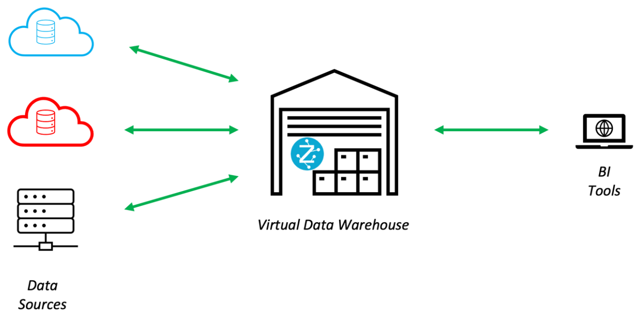


Figure 5: Z-Corp's data landscape with Zetaris: The Networked Data Platform

Zetaris: The Networked Data Platform enables Z-Corp to combine and interrogate its data without moving it.



A virtual schema is a dynamically created logical configuration of all the entities across disparate data sources, enabling them to be manipulated as if they were physically consolidated. Zetaris: The Networked Data Platform ingests the metadata from the source systems to create a virtual schema. The result is a Virtual Data Warehouse (VDW). Zetaris' technology allows Z-Corp to reduce occurrences of data egress significantly. Traffic between data sources is minimized because the data itself does not have to be joined for collective interrogation.

4.2 Optimizing analytical workloads

Zetaris: The Networked Data Platform features a heterogeneous query optimizer.
As we have seen, query optimization is a technique employed to reduce the impact of complex queries on compute, storage and memory. The most efficient plan is selected for the query to be executed.

Zetaris: The Networked Data Platform takes query optimization beyond interrogation of a single database to the simultaneous interrogation of multiple databases registered in the Virtual Data Warehouse. The process is both dynamic and automatic. Every query is optimized at run time.

Beyond this, Zetaris: The Networked Data Platform employs a user-controlled caching feature to reduce the drain on resources – and the associated costs – represented by frequently run complex queries. Frequently run queries can be cached, thereby reducing overhead, increasing run speed and, most importantly, bringing down cloud costs.

5. Conclusion

So, what will this mean for Z-Corp's TCO?

The exact figures are impossible to predict, but a reasonable estimate would be a 30% reduction of the annual increase indicated in Figure 3. This would mean that, by Year 5, costs would exceed projections by something over $750K instead of by more than $1.1M.

The effect on Z-Corp's TCO would be a considerable slowing of the rise indicated in Figure 4.

In this example, Zetaris has saved Z-Corp more than $350,000 over four years. Zetaris: The Networked Data Platform identifies the data used by similar applications and creates common query caches within the VDW (or VDWs). By re-directing processing and data access to the Networked Data Platform, enterprises can create a 'drill-through' path to the cloud data. This drastically reduces the load on the cloud data warehouse CPU. Common queries are pulled down to a local cache within the Networked Data Platform, meaning that repeated execution will not inflate consumption costs.

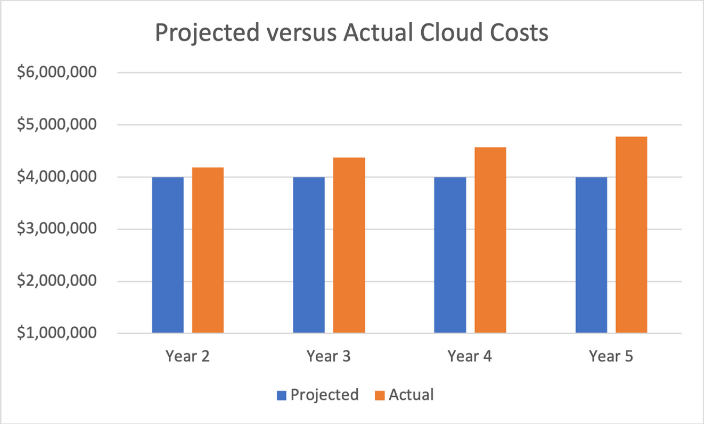


Figure 7: Z-Corp's projected and actual cloud costs with Zetaris

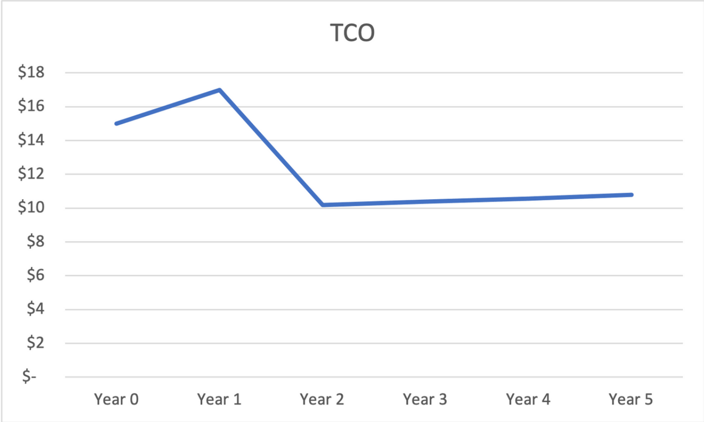


Figure 8: Z-Corp's TCO after movement to the cloud with Zetaris

6. Further Reading

Surveys have found that many companies have left the cloud after enduring much more severe cost blow-outs than the generic example above might indicate. The example is deliberately conservative for the same reason the problem exists: it is almost impossible accurately to predict usage fees.

Discussions relevant to this topic include:

Allen, Tom. ‘NASA forgot about the egress costs for its 247 petabyte data store on AWS.’ https://www.computing.co.uk/news/4012760/nasa-egress-costs-aws.

Han, Yan. ‘Cloud Computing: Case Studies and Total Cost of Ownership.’ https://www.researchgate.net/publication/272658747_Cloud_Computing_Case_Studies_and_Total_Cost_of_Ownership.

Linthicum, David. ‘DevOps in the cloud: 10 steps to success.’ https://techbeacon.com/enterprise-it/devops-cloud-10-steps-success.

McKnight, WIlliam. ‘SQL Transaction Processing, Price-Performance Testing: Microsoft SQL Server Evaluation: Azure Eas-Series Virtual Machines vs. Amazon Web Services R5a-Family EC2 Instances.’ https://clouddamcdnprodep.azureedge.net/gdc/gdcGFnQ2v/original.

Jake Dolezal. ‘Product Profile and Evaluation: Amazon Redshift, Microsoft Azure SQL Data Warehouse, Google BigQuery, and Snowflake Data Warehouse.’ https://gigaom.com/report/data-warehouse-cloud-benchmark/.

Nachmani, Ofir. ‘The Basics of Cloud Capacity.’ https://iamondemand.com/blog/the-basics-of-cloud-capacity/.

Nelson, Lauren and Rich Lane. ‘Top 10 Facts Every Tech Leader Should Know About Cloud Cost Optimization.’ https://www.forrester.com/report/Top+10+Facts+Every+Tech+Leader+Should+Know+About+Cloud+Cost+Optimization/-/E-RES153056.

Orme, James. ‘Size of AWS “data transfer costs” revealed in leaked internal documents.’ https://www.techerati.com/news-hub/size-of-aws-data-transfer-costs-revealed-in-leaked-internal-documents/.

Zibitsker, Boris. ‘Which Platform Is Best for Your Cloud Data Warehouse?’ https://assets.teradata.com/resourceCenter/downloads/AnalystReports/BEZNext-Which-Platform-is-Best-for-your-Cloud-Data-Warehouse-MD008442.pdf.