How to Build a Data Analytics Platform

TO SUPPORT YOUR END-TO-END DATA JOURNEY

MATILLION

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Companies today are increasingly generating large amounts of data to better understand their business, their customers, and their market. Increased data volume alone, however, will not lead to increased success. Instead, companies must find new ways to consolidate and transform disparate data sources in order to generate meaningful insights.

A cloud-based Data Analytics Platform (DAP) can integrate the multitude of data sources within a business into an enterprise-wide analytics system. By unifying the necessary data analytics technologies, a DAP will ingest raw data, transform it, and use it for reporting, analytics, and visualizations - all at scale - helping to quickly draw out relevant insights that inform better decisions.

How well and how quickly your business can put data to work for you is dependent on the capabilities of your DAP. In this whitepaper, we will outline the key tenets of a Data Analytics Platform (DAP) and illustrate how your business can adopt cloud technologies to design a solution that is easy to use, accelerates time to insights, and optimizes your IT spend.

INTRODUCTION

Modern companies tend to generate large amounts of data in an effort to better understand their business. This expanded volume of data, however, does not necessarily provide immediate value. Instead, a business must derive value from its ability to bring disparate sources of data together and transform them into meaningful insights that lead to a better understanding of its market, customers, and business operations.

To accomplish this need, businesses are increasingly adopting data analytic solutions such as Data Analytics as a Service (DAaaS), Insight Platforms, and Data Analytics Platforms. These cloud-based platforms integrate end-to-end data production components into an enterprise-wide analytics system. To get the most out of any DAP platform, the cloud is key to success in this data-driven world. On-premises databases require in-house servers that cannot flexibly scale to match the pace of growing data needs. Alternatively, the cloud offers businesses near-infinite scalability and storage, allowing economies of scale unachievable with an on-premises configuration.

This whitepaper will outline the key tenets of a Data Analytics Platform (DAP) and illustrate how your business can adopt cloud technologies to design a fit-for-purpose solution that is cost efficient and scalable. A DAP can help your business ingest raw data, transform it, and use it for reporting, analytics, and visualizations - all at scale - giving your users the ability to draw out the relevant insights that inform better decisions.

THE PROBLEM

In the past, Business Intelligence (BI) technologies were all housed within on-premises data centers, where local data sources were fed into local data warehouses and reported on by local users. The modern enterprise has grown beyond the limitations of this on-premises and local model. In today's world, an organization might depend on hundreds of different systems, some owned in-house, but more and more supplied to the company as a service from third-party providers. Integrating this variety of data into a traditional data center poses difficulties.

Notably, performance within an on-premises data center is physically limited to what hardware has already been purchased, configured, and maintained by the company itself. Capacity is finite and purchasing room to grow means computing and storage resources are idle until they're needed, leaving bought and paid for resources unused. Once demand approaches existing capacity, more resources need to be purchased, and the cycle continues. For on-premises data centers, this cycle is both time consuming and expensive.

Under this approach, performance suffers until capacity is expanded, causing internal backlogs, preventing customers from using your services, and creating missed deliverables. And, even if companies recognize the need to load data in-house from external sources, network capacity often creates bottlenecks, preventing the movement of modern data volumes into a local data center. Last, the wide varieties of different data sources used today typically require migration through expensive custom coded solutions. Often, leaving behind these data sources is not an option as they represent a critical part of the picture regarding your enterprise's overall activities.

In place of on-premises databases, cloud-based data warehouses remove the physical limitations and maintenance overhead with fully managed solutions possessing near-infinite storage capabilities. The problem of data consolidation and aggregation, however, still persists. Businesses need to develop the capability of not only moving all their data from a wide variety of sources into a cloud-based data warehouse, but also coordinating those complex, interdependent activities. Once loaded, data needs to be cleansed, enriched, aggregated, and otherwise transformed to generate meaningful insights. Finally, the resulting datasets require presentation in accessible ways so that the users who depend on them can understand the outputs and incorporate them into decision-making processes, leading to better outcomes.

THE SOLUTION

All of the challenges described above are solvable. Creating a cloud-based **Data Analytics Platform (DAP)** can remove the obstacles outlined previously and not only streamline your data workflows but also empower your business to effectively manage its complete end-to-end data analytics journey. When it comes to technology facilitating this journey, your business' biggest investment will most likely be its cloud data warehouse. Focusing on getting the most value out of that cloud data warehouse should be paramount. Luckily, many of the industry-leading cloud data warehouses, such as Amazon Redshift, Google BigQuery, and Snowflake, have substantial partner ecosystems that provide compatible third-party solutions that work with your existing cloud infrastructure.

In addition to selecting a cloud data warehouse, you should also focus on adopting the right technologies for your business needs instead of settling for an expensive, all-in-one solution that contains features beyond your scope. Doing so allows you to create an affordable but flexible platform best suited for your business data. Using the notion of a DAP can help bring together your solution including technologies for Analytics, Business Intelligence, as well as a Cloud Data Warehouse (CDW). Uniting these technologies together will create a DAP that not only meets your current requirements but provides an extensible solution that can be modified and expanded to meet ever-changing and complex use cases. For example, such a framework allows you to bring in Data Lakes, Data Vaults, Machine Learning, and Artificial Intelligence to further bolster your data analytics capabilities. Agility is key as such platforms are developed over time (not overnight) according to the most pressing business needs.

How well and how quickly your business can put data to work for you is dependent on the capabilities of your DAP. In addition to the technologies mentioned above, another fundamental component of your DAP is data integration and transformation. Since your DAP aims to support your end-to-end data journey, it needs to be able to handle raw, diverse data forms to turn them into clean datasets for analysis, business intelligence, and visualization. This last piece is often referred to as an ETL (extract, transform, and load) solution to support moving and transforming your data. ETL solutions vary however in their approach and scope, a topic that we will discuss later in this document.

Using Matillion, the data transformation solution for cloud data warehouses, as part of your DAP will allow you to leverage the power of your cloud data warehouse and optimize the performance of your data visualization tools. Learn more about how Matillion customers are improving their data processing power and reducing IT spend.

THE BENEFITS

Once you have your Data Analytics Platform technologies in place (Cloud Data Warehouse, ETL solution, data visualization layer, etc.), you are now ready to realize the benefits such a platform provides. Here are some of those benefits:

Extensibility

Integrate new technologies and adaptable components to simplify your business' end-to-end data journey

Simplicity

Simplify your end-to-end data journey with flexible technologies that apply and scale to multiple use cases

ROI

Get more value from your IT spend with pay-as-you-go technologies that work together to ingest, transform and analyz data

Speed

Cloud technologies with intuitive user experiences reduce development time and speed up the release of actionable analytics over hand coding

Agility

As you learn from your data and development, patterns you can adaptively add components to the platform incrementally incorporating new data sources, transformations, and analyses

Scalability

Whether you are dealing with a handful or hundreds of data sources, a DAP can meet your current and future needs by offering an enterprise scale solution to data analytics

WHEN TO USE A DATA ANALYTICS PLATFORM

The difference between a common Business Intelligence (BI) technology stack and a Data Analytics Platform is primarily a difference in scale. Where the former may feed one or a few source systems into a data warehouse and then report on the same, a DAP is used to connect to dozens or hundreds of data sources, load those raw sources into a data acquisition staging area, then transform that raw data into any number of data warehouses and marts. This curated data ultimately feeds reports, visualizations, and analytics for a wide variety of users, from power-users who perform their own advanced, statistical analyses on raw data to end-users who analyze highly processed data for their particular business unit using Excel. The resulting analyses ultimately feed into a data-driven decision making process.

While the general pattern here may be the same as a traditional BI stack, the scale is radically different. The DAP represents a one-stop shop for an entire company's data delivery needs and provides a 360-degree view of your enterprise's activities. You should consider building a DAP when you need a high-level view of all your corporate activities and doing so requires synthesizing data from numerous systems: marketing, sales, finance, logistics, etc. These disparate sources might be available in wildly different formats, accessible from relational databases, RESTful APIs, or file-based artifacts on distributed file systems. A DAP allows you to corral such sources in a centralized and cloud-based data warehouse, manipulate the data to meet reporting requirements, and ultimately visualize and analyze enterprise activity to inform decision making.

Matillion transforms your data to optimize the power of your business intelligence (BI) applications. Matillion's pre-built connectors will help you not only consolidate all your data sources (applications, data lakes, databases, etc.) but also perform data transformations that can be easily interpreted by data visualization and analytic tools to drive better business decision making.

USING A DATA ANALYTICS PLATFORM EXAMPLE: GLOBAL MEDIA AND ADVERTISING

As a potential DAP use case, consider a hypothetical global media and advertising company. Their centralized IT group, charged with providing data to their satellite offices for the purposes of planning media purchases, opts to build a Data Analytics Platform to accommodate hundreds of different data sources and a wide variety of user needs. The DAP allows them to do away with dozens of local data centers hosting idiosyncratic, one-off processes in favor of their Cloud Data Warehouse, affording the company limitless scalability as well as a consistent, efficient, and company-wide view of the data.

The ultimate goal of this platform is to produce data-driven decisions. In this case, where and when to place which advertisements and on what medium including search, social media, or television advertisement campaigns. The DAP provides for the entire company's data analytics needs, offering automated client deliverables as well as live, interactive dashboards sourced from a cloud-based, cross-channel repository that also supports data science models and simulations.

With user needs ranging from raw data to dashboarded reports, the central IT group opts to use Amazon Redshift as their database engine. Amazon Redshift's easy scalability makes it a preferred option along with its Massively Parallel Processing (MPP) capabilities and columnar orientation for handling huge volumes of data. On top of performance benefits, Amazon Redshift offers a pay-as-you-go model to alleviate the need to purchase unused capacity, once a fundamental and cost-prohibitive roadblock preventing the creation of such platforms in local data centers.

To integrate their hundreds of myriad source systems into Amazon Redshift, the media company chooses Matillion due to its ease of development and numerous, out-of-the-box source system connectors. Matillion also employs an extract, load, transform (ELT) architecture which leverages Amazon Redshift to its fullest by applying the CDW's near-infinite resources to the load and transformation processes. Matillion also fits into the same pay-as-you-go economics model as Amazon Redshift itself, thereby limiting costs while you grow and develop your DAP.

The analytics and visualization layer in this example DAP employs Tableau for it's easy dashboarding as well as ad-hoc reporting capabilities. Power users such as Data Scientists may even pull raw data from the staging layer via Python or R. Users can even connect their Excel workbooks to the cleansed data, providing a single source of truth and avoiding the data siloing and fragmentation historically associated with this practice.

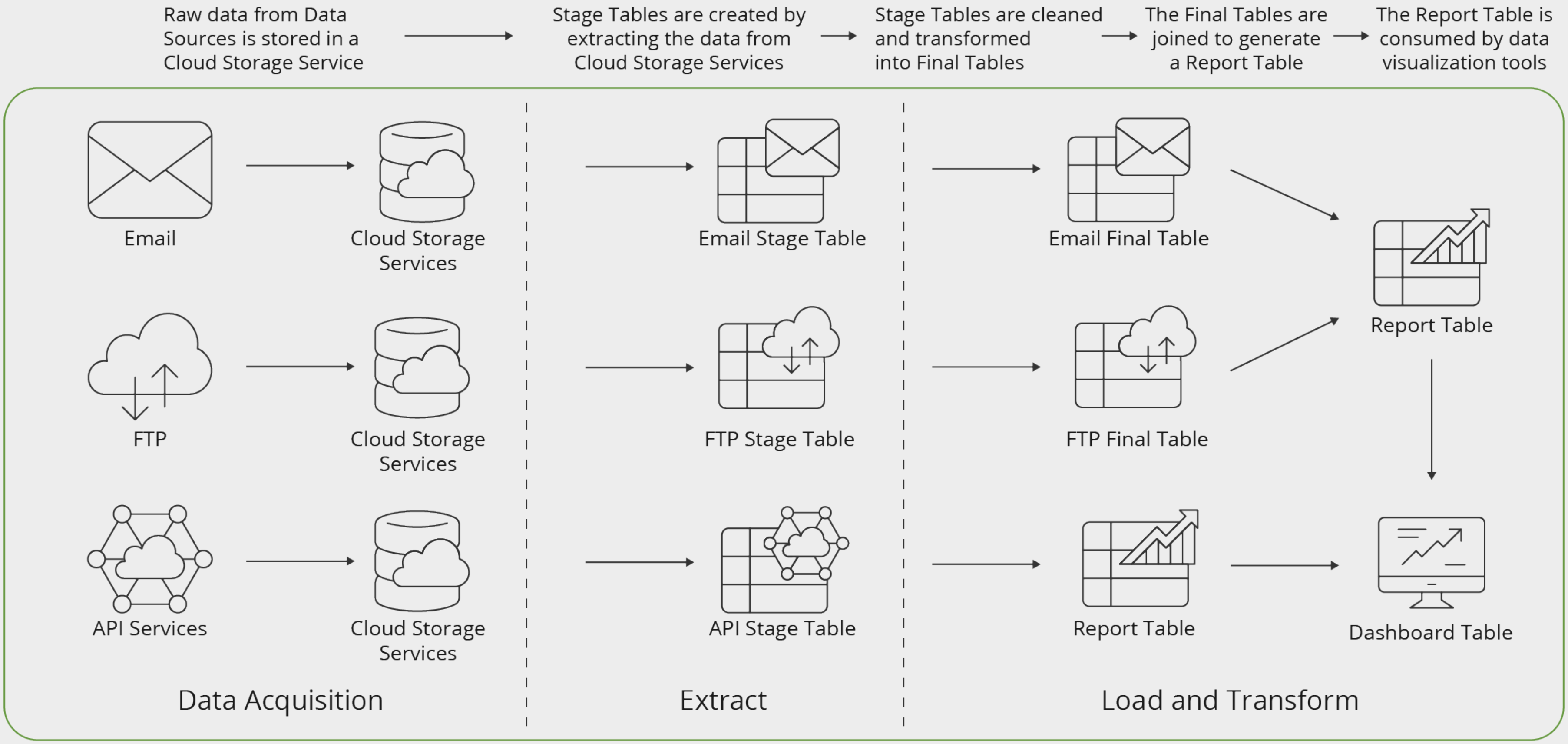

Raw data from Data Sources is stored in a Cloud Storage Service
Stage Tables are created by extracting the data from Cloud Storage Services
Stage Tables are cleaned and transformed into Final Tables
The Final Tables are joined to generate a Report Table
The Report Table is consumed by data visualization tools
Email
Cloud Storage Services
Email Stage Table
Email Final Table
Report Table
FTP
Cloud Storage Services
FTP Stage Table
FTP Final Table
API Services
Cloud Storage Services
API Stage Table
Report Table
Dashboard Table
Data Acquisition
Extract
Load and Transform

Diagram 1: Data Analytics Platform example - Global Media and Advertising

Though the diagram above represents a small, simplified slice of the activities within this DAP, it nevertheless demonstrates the principle data flows provided by such a platform. Here we see data extracted from numerous source systems including email attachments, as well as vendor supplied FTP and API services. Those sources are staged into Redshift, cleansed and aggregated into consistent data structures for reporting through Excel extracts and Tableau dashboard. The DAP also provides our Data Scientists with the raw data they need to model advertisement effectiveness, all of which feeds into data-driven advertisement placement decisions.

HOW TO BUILD A DATA ANALYTICS PLATFORM

The Technologies

Analytics: Working backwards, the final layer in a DAP performs data analysis and visualization. This layer provides your analysts with high quality information related to your enterprise's activities, allowing optimal, data-driven decisions to be made by your executives. Modern tools in this space include Tableau and Looker (among others) and provide interfaces for ad-hoc analysis as well as enterprise-class business intelligence reporting. Other tools in this layer may range from simple Excel to the more robust RStudio. Such tools leverage the data transformed and curated in your cloud data warehouse by your data transformation solution and extracted from upstream source systems.

Business Intelligence: Sometimes your business intelligence technology may be the same as your analytics or visualization technologies, or even both. What is important here is to make sure you have the capability to use your data to drive intelligence. That is, you have a way of using data to formulate business insights and decisions, whether that be data mining, dashboards, building reports, or creating visuals. Looker, Tableau, AWS Quicksight, Data and Studio and PowerBI are popular options.

Cloud Data Warehouse: Cloud data warehouses are cloud-native database platforms specifically designed for easy (and even automated) scalability as well as massively parallel processing capabilities. Often column-oriented for optimized I/O operations, these platforms offer near-limitless compute and storage capabilities. Modern cloud-data warehouses include Amazon Redshift, Google BigQuery, and Snowflake.

Data Integration and Transformation: To create a DAP, your business needs to integrate a wide variety of relational, structured, and semi-structured data sources. These data sources need to be manipulated and transformed into a unified format that feed downstream analytics and visualizations. A cloud-native data transformation solution allows high-speed connections between your enterprise and other cloud-based data sources. Matillion is an example of a data transformation solution that can help you extract data from your on-premises or cloud-based applications, move that data into a cloud data warehouse, and then transform data into the unified formats necessary for your BI and analytics tools.

Matillion provides the capability required to manipulate your business data into the right format for your business intelligence (BI) application. Consolidate data from source systems, data lakes, on-premises databases, and other cloud data warehouses for advanced analytics using pre-built connectors. Then perform transformations to produce production quality data that can be interpreted by visualization and analytic tools, to inform your business' decision making.

The Architecture

Multiple and fragmented data analytic solutions can create competing versions of the truth within an organization. A DAP forms a complete, end-to-end data solution offering enterprise-wide data governance as well as a unified corporate view. Such solutions provide not just ingestion but also data transformation according to established business rules, preparing data for downstream use by analysts and decision makers.

Data Analytics Platforms (DAPs) combine complementary technologies to create a unified system that:

- Ingests data from a variety of sources such as data lakes, databases, APIs, and files in numerous formats
- Transforms the ingested data by joining it to other data sources, as well as cleansing, aggregating, or otherwise manipulating it
- Visualizes the transformed data graphically on dashboards or reports
- Forms the basis for executing well-informed business decisions.

The diagram on the next page depicts the high-level components that might comprise a modern DAP. Extract and load processes move data from numerous and disparate data sources into the Cloud Data Warehouse (CDW), ideally and initially into a section of the same reserved for raw, staging data. From there, the near-infinite computing resources available to such CDWs can merge the staged data into an Operational Data Store (ODS), often structured to match the layout of the sources themselves.

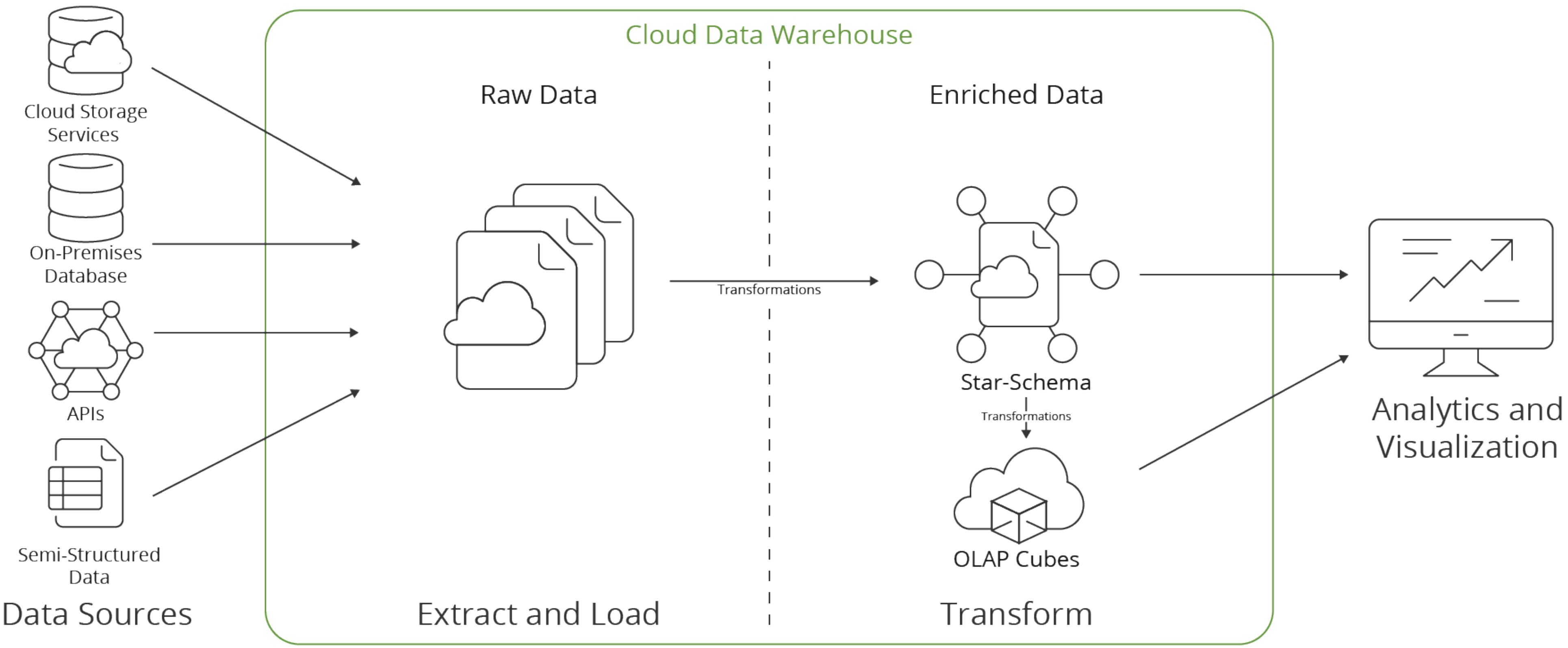


Diagram 2: Architecture Diagram - Data Analytics Platform

Those same compute resources can then transform the often highly-normalized ODS structure into one more suitable to reporting purposes, frequently duplicating and de-normalizing data into a classic star-schema. That star might be further optimized for reports through roll-up transformations that create OLAP cubes. Finally, these more performant data architectures might then be consumed by the DAP's visualization layer, offering end users (such as analysts or executives) insight into enterprise-wide activities that inform the best decision making possible.

The section below expands on each of these DAP components in turn, starting with various Data Sources, an Integration and Transformation Layer, a Cloud Data Warehouse, and the Visualization and Analysis Layer.

Data Sources

The items below represent a few common data sources in today's enterprise environments:

On-Premises Databases: Many organizations struggle to transition completely into cloud platforms and still maintain databases in on-premises data centers. Luckily, incremental, hybrid approaches are perfectly feasible, allowing you to keep running existing databases locally while strategically extracting their critical data elements and loading the same into a CDW.

Data Lakes: Defined by storing files in native formats, flexible access, schema-on-read, as well as the decoupling of storage from compute resources, data lake architectures represent more and more common repositories for enterprise activity. With near-limitless extensibility, cloud-enabled storage platforms such as AWS S3, Azure Blob Storage, and Google Cloud Platform Cloud Storage often provide the data lake's persistence layer. Loading files into your company's CDW from these platforms represents critically important functionality.

Semi-Structured Data: Related to data lakes, which are often composed of semi-structured files, modern DAPs often need the ability to load XML, JSON, or even spreadsheet files. Choose tool sets that can load CDWs from these files.

APIs: Application Programming Interfaces allow integrations between your company and important, third-party resources like Salesforce or Google Analytics. As such, critical company data may be stored within these Software-as-a-Service (Saas) platforms; extracting data from these sources frequently provides important insights. Use tools that support the incorporation of these data sources into your CDW.

Integration and Transformation Layer

Often referred to as ETL packages for the Extract, Transformation, and Load operations such software provides, DAPs rely heavily on products that can integrate the data sources listed above, among others. Doing so requires not only the ability to extract and load data from a wide variety of platforms and get that information into the CDW, but also manipulate the resulting data into structures best suited for gleaning insights. Typically, ETL software runs on hefty dedicated servers with sizable compute, memory, and storage resources, both to provide space to land extracts from remote source systems as well as transform those extracts into the format required by the use case. Once transformed, the results are loaded into the target. As discussed above, scaling dedicated physical hardware is a time consuming and expensive prospect; the same applies to ETL infrastructure which creates significant performance bottlenecks when operating beyond capacity.

To avoid the risk of poorly performing hardware and the cost of upgrading ETL infrastructure, more and more professionals employ an ELT (Extract, Load, Transformation) approach, flipping the usual order of events so that data is extracted and loaded first, and then transformed inside the CDW. Solutions that invoke an ELT method instruct the CDW to use its scalable and massively parallel processing capabilities to load your data into the cluster, then transforming it into the form you specify once that load completes. Modest computing capabilities suffice to generate these instructions and allow you to leverage the investment you've already made in your cloud data warehouse platform, which handles the heavy lifting required during the ELT process.

The ELT approach wouldn't be possible without the innate scalability built into CDW platforms, which provides the resources needed for these operations and allows business intelligence organizations to forgo the more costly and less flexible ETL pattern. Plus ETL software is extremely expensive to license. These licenses often require extra fees for higher data volumes and charge even more for connector add-ons that give you access to the source systems you need to load into your warehouse. The ELT approach provides real benefits, allowing your organization to push its integration and transformation workloads into the platform best suited for it, namely the CDW.

Matillion's cloud-native data transformation architecture leverages the virtually infinite storage and compute resources built into modern cloud data warehouse platforms. By using Matillion and a cloud data warehouse, you are able to rapidly scale your data transformation capabilities by taking advantage of the power and features of the cloud.

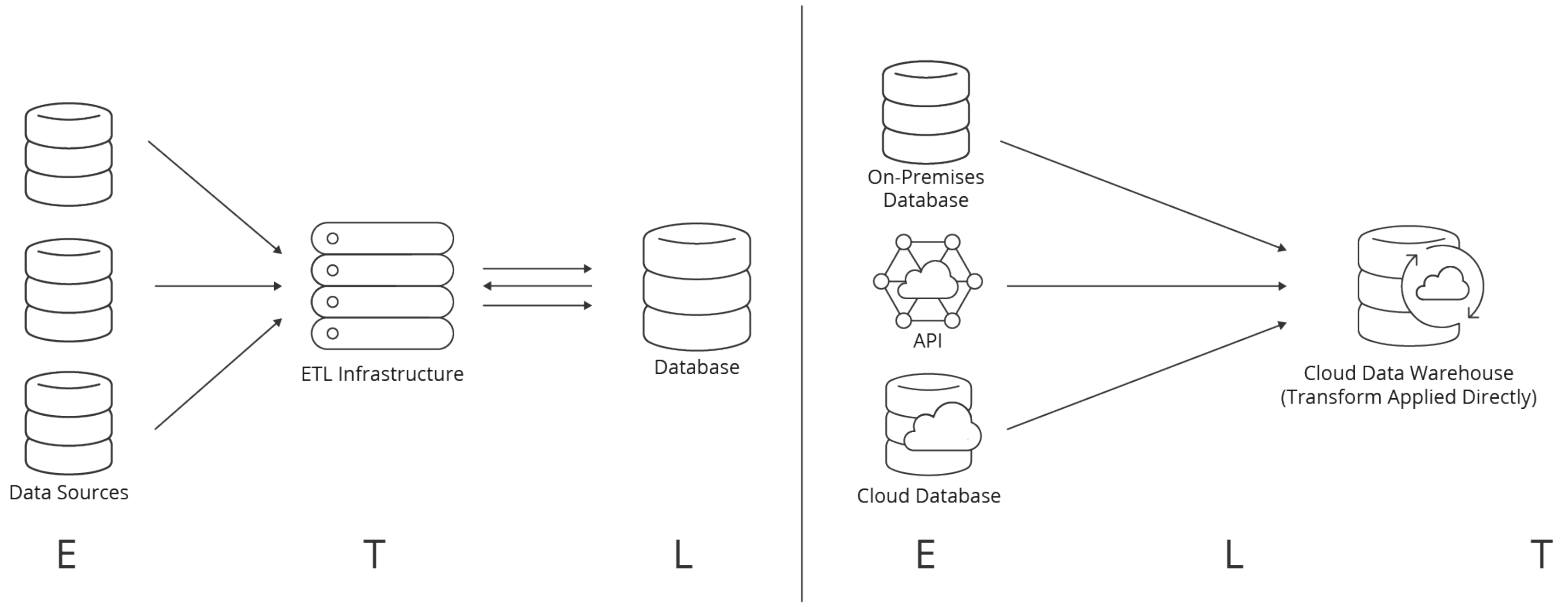


Diagram 3: Extract, Transform, Load (ETL) vs. Extract, Load, Transform (ELT)

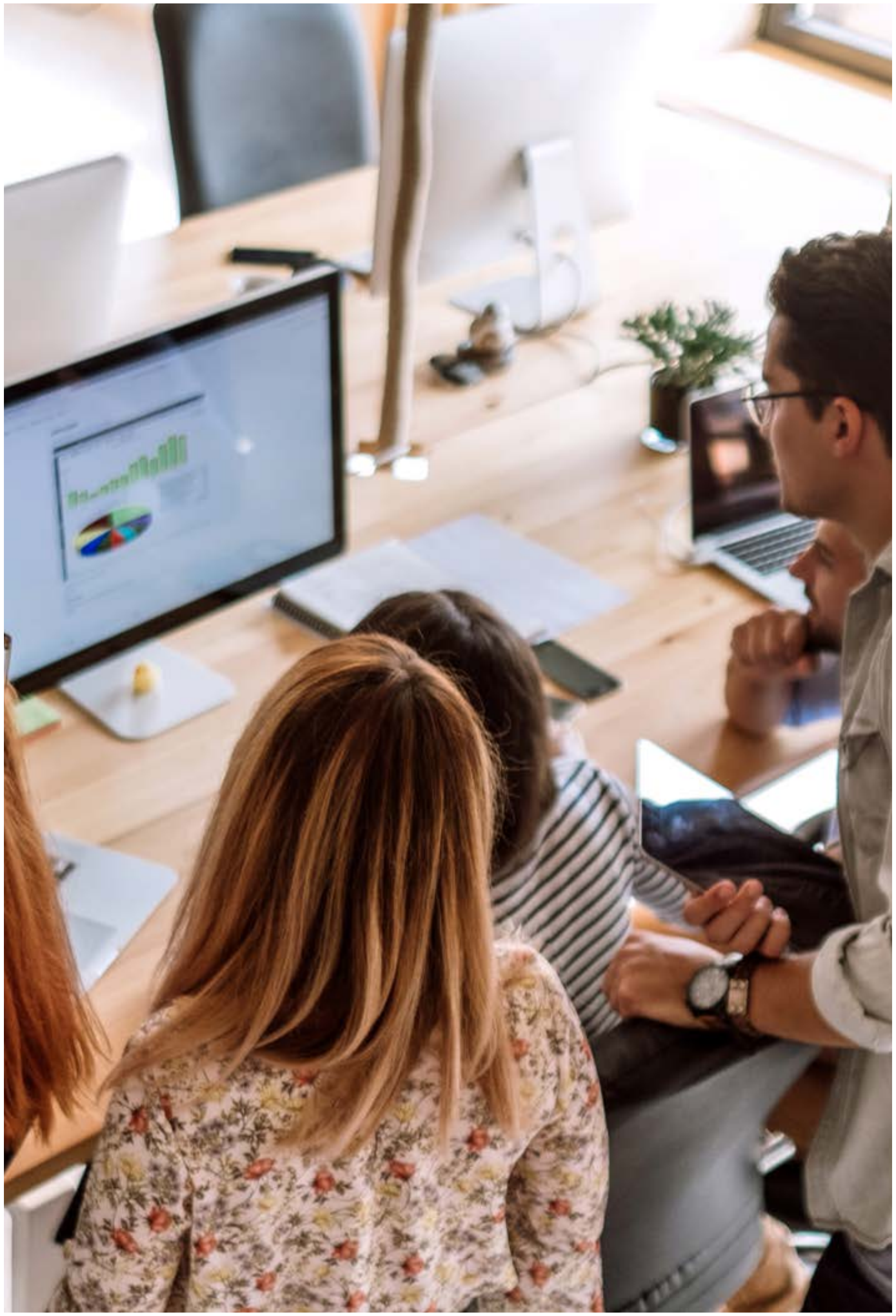
Matillion's cloud-native data transformation architecture is especially helpful here in that it leverages the virtually infinite storage and compute resources built into modern cloud data warehouse platforms; your company can leverage the investment they've already made here to transform data after it's loaded as-is from the source. Notably, Matillion's pay-as-you-go pricing model lets you apply the principles of cloud-economics to data transformation. Our hourly price includes any data source connector Matillion offers, as well as free support and a simple but powerful GUI interface for transforming your data inside your cloud data warehouse.

Visualization and Analysis

The final component in the DAP is a visualization and analytics layer. This layer provides both formal and ad-hoc reporting capabilities based on the final, transformed data we've extracted from our source systems, and enriched with other data elements to create actionable business intelligence. The resulting visualizations and analyses inform the data driven decision making process, aiding better decisions and ultimately generating competitive advantages.

Tools like Tableau, Looker, AWS Quicksight, and Google Data Studio currently provide best-of-breed visualization and reporting suites. Such packages give your organization interactive reporting and dashboarding feature sets that elegantly depict insights not easily gained from spreadsheets alone. Their GUIs are both easy enough to rapidly start producing meaningful information and rich enough to provide power users with all the advanced features they require. Plus these tools support CDWs allowing your enterprise to combine these flexible, scalable, and cloud-native components into an extensible DAP framework.

More technical and statistically savvy audiences may use RStudio or Python Notebooks to create advanced data models, or even leverage Artificial Intelligence (AI) services such as Google Cloud Platform's Cloud AI or AWS' SageMaker. At the same time, many users are familiar and comfortable with Excel as an analysis tool. The analysis and visualization layer can include many technologies, all being fed from the curated data maintained in the CDW.



GETTING STARTED

We have outlined many of the technologies and requirements to create a new Data Analytics Platform that will help you improve your business insights as well as help you understand and connect with your customers, innovate faster, and streamline your employees' workloads.

While there are various approaches to creating a DAP, we recommend the following approach in order to maximize time to value, leverage the flexibility of the cloud, take advantage of compute and storage resources, and minimize costs. Following this approach will lower the barriers to entry to acquire an advanced data analytics platform for your business.

Following this approach will allow you to quickly and efficiently get your DAP up and running and allow you to better harness the power of your data to help drive your decision making.

Steps to Get Started

1. Select the cloud based data warehouse that meets your business 'current and future data needs. At the time of writing this whitepaper, current market leaders include Amazon Redshift, Google BigQuery, or Snowflake.

2. Identify the data sources and applications that your business will need to ingest into your cloud data warehouse. If your short list options use licenses at a fee or charge for connects be sure to include current and projects costs in your expected total cost of ownership.

3. Select a data transformation solution. We would recommend considering a data transformation solution that also handles data loads. This will help you streamline your data journey and simplify your DAP. Where able, find solutions that reduce your need to hand code to promote resiliency.

4. Select a business intelligence and analytics tool, if you haven't already. Current market leaders include Looker, Tableau, AWS Quicksight and Data Studio.

5. Create a proof of concept (PoC) engagement to test your data migration and transformation strategy as well as the quality of your data once complete.

6. Execute your data migration and transformation strategy.

7. Connect your BI tool to your data warehouse, etc.

Speed. Savings. Simplicity. Scalability.

Put Matillion at the center of your Data Analytics Platform to orchestrate your end-to-end data journey.

- ✓ An instant-on purchase experience via the AWS Marketplace, Google Cloud Marketplace and Azure Marketplace to get you up and running in just a few clicks
- ✓ Includes 60+ native data connectors out-of-the-box
- ✓ Pay-as-you-go billing model, to eliminate lengthy procurement processes
- ✓ An intuitive user interface, to minimize technical pain and speed up time-to-results
- ✓ Transform data to create production-quality datasets for visualization and analysis

GET A DEMO

Get a demo to see first-hand how Matillion fits into your Data Analytics Platform.
www.matillion.com/get-a-demo

MATILLION

CONCLUSION

We hope you learned how a cloud-based DAP can help your business own the end-to-end data journey. Implementing a unified data analytics technology stack that is fit-for-purpose, simple to set up, and easy to use will help your business save resources by generating more value for your IT spend, speed up time to insights and scale to meet future data growth and needs. This way you can start putting your data to work for you by generating relevant insights that aid better decision making and help drive competitive differentiation within your market.

About Matillion

Matillion is an industry-leading data transformation solution for cloud data warehouses. Delivering a true end-to-end data transformation solution (not just data prep or movement from one location to another), Matillion provides an instant-on experience to get you up and running in just a few clicks, a pay-as-you-go billing model to cut out lengthy procurement processes, and an intuitive user interface to minimize technical pain and speed up time to results. Matillion is available globally for Amazon Redshift, Snowflake, and Google BigQuery on leading cloud infrastructures.

Find out more at www.matillion.com.