MATILLION | Google Big Query

Delivering **business-ready** data at **speed and scale.**

Realize the value of Google BigQuery faster, with Matillion.

Start

Data, **data everywhere!**

As “software eats the world[1]”, over 25,000 SaaS companies now exist[2], from CRM and ERP to HR and Customer Service to Cloud Data Warehouses and Data Lakes ... the list goes on and on.

If all these systems had standardized on the same data formats, and had clean and complete data, it would be much easier to unify data from them for analytics and operational use cases. Unfortunately that is not the case.

The reality is that companies have on average **400 disparate data sources** with varying formats, the amount of data businesses need to manage is growing **63% per month**[3], and McKinsey predicts that by 2025 data will be **embedded in every decision, interaction and process**[4].

As a data practitioner, you need to work through the challenge of keeping up with a constantly accelerating growth of distributed and diverse data sources so the business can get the most from all of its data.

[1]Source: https://a16z.com/2011/08/20/why-software-is-eating-the-world/
[2]Source: https://ascendixtech.com/number-saas-companies-statistics
[3]Source: https://www.matillion.com/resources/blog/matillion-and-idg-survey-data-growth-is-real-and-3-other-key-findings
[4]Source: https://www.mckinsey.com/business-functions/quantumblack/our-insights/the-data-driven-enterprise-of-2025

Go from zero to **business-ready data.**

As companies grow they use more applications and data sources. As they become more competitive, there is a more urgent need to use their data to be more customer responsive, and make better and faster decisions.

With increased users and systems that need to use data across the business, the data landscape matures from data silos, to data centralization, to data ready for the business. As organizations move from one phase to the next, they will reap increasing benefits from their data in a number of areas.

Data Centralization
(Centralized data, but not clean and complete. Lots of prep work for the business)



Data Silos
(Data, data everywhere. Hard to find and use)

Data ready for business
(Timely, clean, complete data for business)

The journey from silos to **business-ready data.**

	Data Silos	Data Centralization	Data Ready for the Business
Data state	Data is not centralized or shared.	Data centralized into Google BigQuery and shared for analytics, data science and operational use cases.	Data centralized into Google BigQuery and shared for analytics, data science and operational use cases.
Data quality	Inconsistency across each app based on data entry rules.	Some standards set. Data end users (e.g. analysts) spend additional time prepping and cleaning data.	Company standards set. Data is clean, complete and deduped in the cloud data platform.
Data synchronization	Lack of synchronization when changes are made.	Changes in Google BigQuery often not synced back to operational systems, leading to multiple versions of the truth.	Changes in Google BigQuery synced back to operational systems for consistency.
Data Accessibility	Limited. Hard to find and use data by non-data owners.	Better, but still siloed e.g. the marketing cloud data platform is just used by marketing.	Governance and processes in place. Data catalog with lineage used. Analysts and end-users can find and access data.
Productivity	One-off integrations mean starting from scratch each time.	Lots of coding and duplicate data pipelines across the business slows response time. High costs to monitor and maintain code.	Graphical tools and APIs to automate data pipelines, leading to resource efficiencies and quicker response times.
Decision-making	Based on "gut-feel" since not all data is used in decisions.	Better decisions, but slow to get started and respond to changing data requests.	Better and faster decisions are made through automation, and clean, complete data.
Example analytics use case	Salesforce prospect data analyzed, but lack combined insight of website activity.	Salesforce, Marketo, Google Analytics data loaded into cloud data platform. Hand-coding efforts for validating, deduping and transforming data means slow response times and difficulty scaling.	Governance established. Salesforce, Marketo and Google Analytics data loaded into cloud data platform. Automation of data pipelines to create clean and complete data with timely delivery.

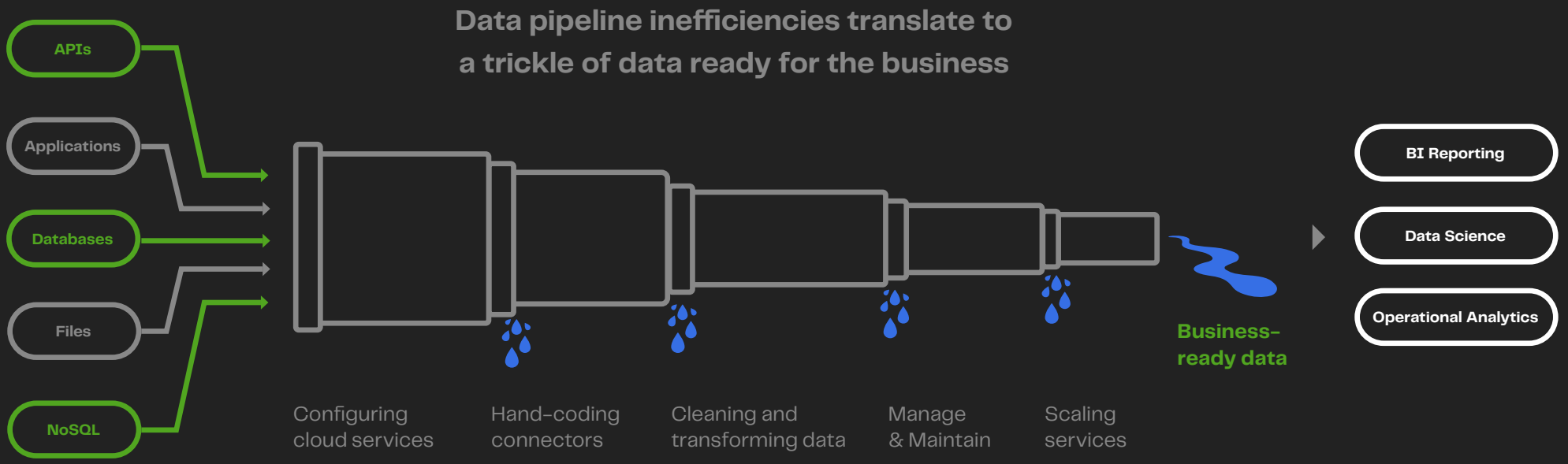
Unfortunately, many companies move to data centralization, but then fall short of delivering at the speed and scale that the business needs. Global data company Quantexa has found that more than 95% of organizations are unable to bring together all their data for effective decision making[5].

[5]Source: https://www.prnewswire.com/news-releases/data-decision-gap-risks-holding-back-economic-recovery-global-research-finds-301419395.html

Challenges delivering **data at speed and scale.**

Speed is paramount to making informed decisions and having a competitive advantage. If you are not acting fast enough, you are making decisions based on information that is inaccurate and out of date. And that will negatively impact your operations.

However there are many potholes and speed bumps that hinder your productivity on the road to fast, clean, complete data.



Challenges.

1. **Managing increasingly distributed and dynamic data.** The amount of data that needs to be managed continues to grow faster than ever. On average, data volumes are growing at a rate of **63% a month**[6], and increasingly you need to integrate both batch and real-time updates.

2. **Getting data into the right format.** Data from SaaS and on-prem apps in structured, semi-structured, and unstructured formats needs to be unified. In a recent survey of 200 enterprise data professionals, **over 90%** said it is challenging to make data available in a format usable for analytics[7].

3. **Hand-coding and using legacy integration tools.** Building an API yourself averages 4 to 6 weeks[8], and using legacy tools not designed for the speed and agility of the cloud will slow down response times and provide subpar performance.

4. **Maintaining data pipelines.** For some businesses, code maintenance takes **30% of the average developer's week**[9]. Too much time is often spent on pipeline maintenance and troubleshooting issues instead of implementing new innovative services.

5. **Hiring and retaining data professionals.** It is very hard to find and retain resources like data engineers, data scientists and Apache Spark developers. According to Payscale, Spark proficient data engineers cost **10-15% more** than general data engineers[10].

6. **Finding and getting access to the data.** IT does not own and control all the systems anymore. Hundreds of apps and data stores exist across the business e.g. marketing, finance, HR, services, manufacturing and partners.

7. **Mastering many technologies.** Data engineers need to be proficient in languages like Python and SQL, tools like Apache Spark and Kafka, cloud data platforms like Google BigQuery, analytics tools like Tableau and Thoughtspot, and data storage, processing and messaging tools for Google Cloud Platform.

8. **Learning paradigm shift from on-prem to Cloud data integration.** Related to #7, there is an architecture and best practices learning curve the data team goes through, e.g. batch to real-time processing; ETL to ELT; implementing cloud security and scaling; governance; and managing consumption-pricing.

9. **Implementing data quality.** Using incomplete and inaccurate data will lead to less than optimal decisions. Cleansing and prepping data can take up to **80% of an analyst's time**[11].

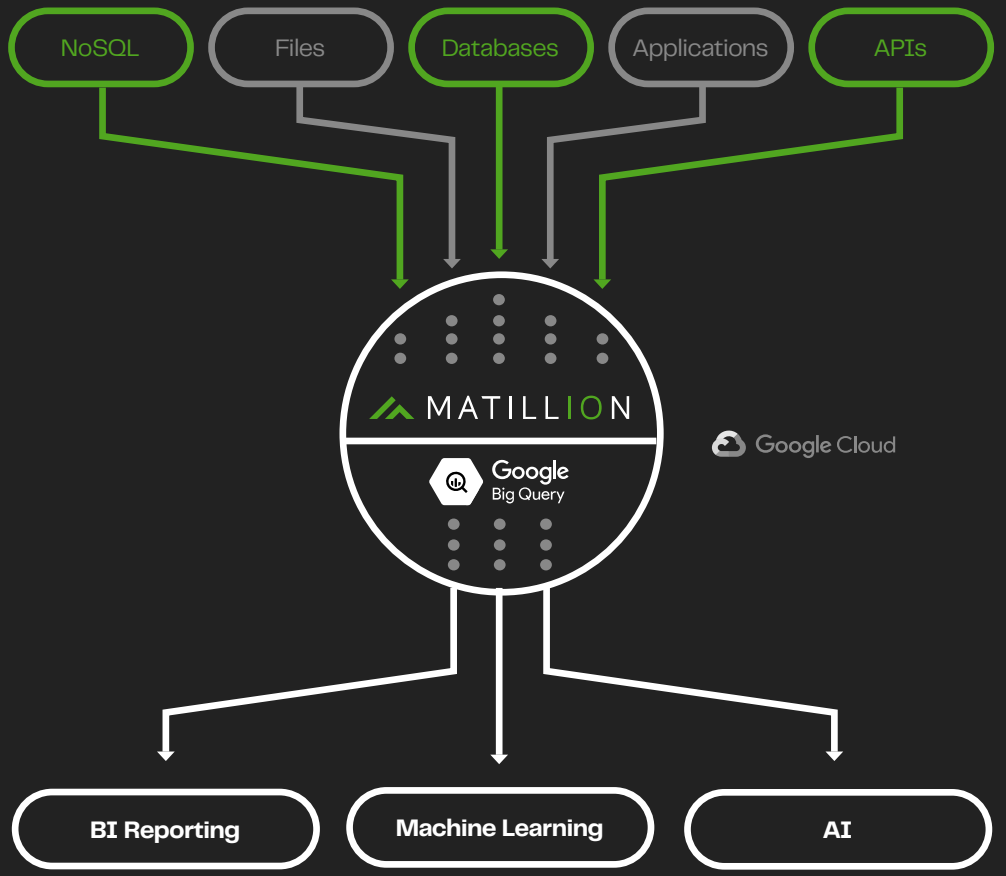
10. **Data monitoring, auditing and compliance.** Many tasks, which are tedious without automation, need to be completed to lower business risks, including: complying with regulations like HIPAA, GDPR and SOC 2; masking private data such as card numbers; implementing role-based access control (RBAC) to keep data secure; and providing an audit trail for compliance.

11. **Handling an ever increasing amount of requests from data analysts and data scientists.** With more data usage comes more data questions that can lead to better company performance. If you cannot keep up with requests, you cannot maximize business potential.

[6]Source: https://www.matillion.com/resources/blog/matillion-and-idg-survey-data-growth-is-real-and-3-other-key-findings
[7]Source: Ibid
[8]Source: https://medium.com/yourapi/how- much-does-it-cost-to-build-an-api-925b1bf90da9
[9]Source: https://blog.tidelift.com/developers-spend-30-of-their-time-on-code-maintenance-our-latest-survey-results-part-3
[10]Source: https://www.payscale.com/research/US/Job=Data_Engineer/Salary/f6851241/Apache-Spark
[11]Source: HBR May-June 2017; HBR Sept 2017

Accelerate reliable **data insights.**

Matillion ETL for Google BigQuery helps data teams get things done faster in the cloud. Deliver business-ready data at speed and scale using a low-code/no-code platform to load, transform, orchestrate, and sync your data. Matillion ETL integrates BigQuery-specific functionality and best practices directly into data jobs for improved productivity and performance.



ThoughtSpot

tableau

Power BI

Leverage your **cloud data at scale, with Matillion.**

Easy data loading. Intuitive user interface with ability to leverage pre-built connectors or create new connectors to virtually any source in minutes. Connect to legacy or SaaS, from CDPs to apps. Support for CDC and batch.

Transform data at speed. Leverage Matillion's low-code/ no-code platform for building efficient, high-volume transformations (ELT) in Google BigQuery. Validate, enrich and cleanse data to make it accurate and complete.

Built for scale. Optimized integration and processing on Google BigQuery including:

- Native push-down integration
- Support for materialized views
- Support for semi-structured data including Flatten Structs (nested fields) and Arrays (repeated fields)
- Easily integrate GCP Pub/Sub alerts & notifications

Orchestrate data flows. From ingestion to consumption, with consistent metadata regardless of underlying infrastructure.

High-end security capabilities. Provides advanced deployment options and rights management, leveraging governance.

Easy to maintain. Configure changes to jobs using a graphical UI vs coding, plus job reuse lets you respond faster to change requests.

Sync-back data. Sync processed data back to operational systems, maintaining a single version of the truth.

Easy to manage. Schedule, monitor and manage all pipelines in one place. Automate tedious tasks and ensure compliance.

Simple PAYGO pricing. That gets even more affordable as you grow. Consolidate billing on your GCP account.

Realize value fast. Rapidly progress from proof-of-concept to production to ROI.

Get your free trial at
www.matillion.com/trial



MATILLION | Google Big Query

Back