Go Green Software

A guide for Data Scientists

Google Cloud

Contents

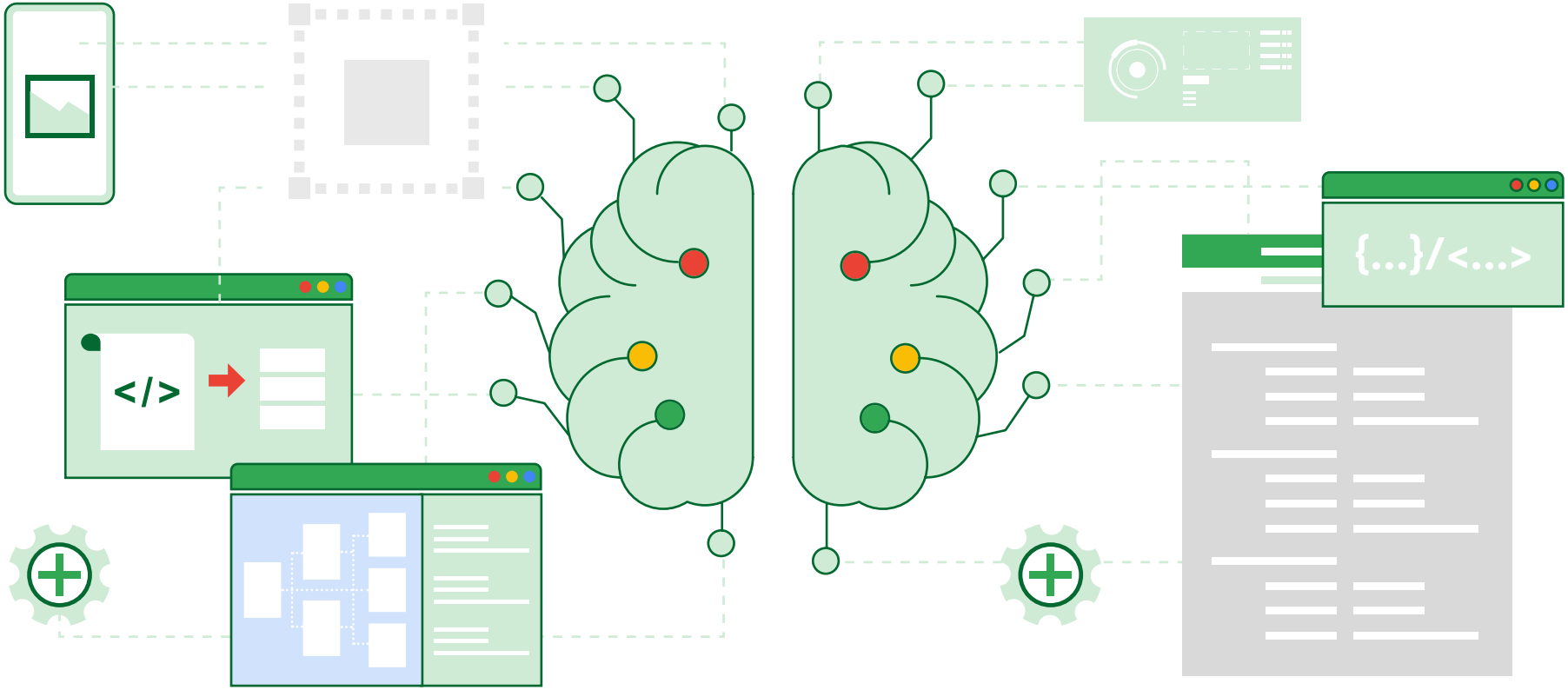
Introduction

Data science and machine learning (ML) are becoming core capabilities for solving complex real-world problems, transforming industries, and delivering value in all domains.

To apply ML effectively, organizations need large datasets, large amounts of on-demand compute resources, and specialized accelerators for ML on various cloud platforms.

These capabilities are complemented by the skills of data scientists, who have the expertise to build complex ML models and pipelines. There are many ways that ML can transform businesses, and each industry has its own unique opportunities.

However, the large amounts of compute and storage resources required for ML can also lead to CO_2 emissions from storing data and running ML models. Data scientist should be aware of these resource-intensive aspects and plan for an optimized usage. The following practices can help to minimize the carbon footprint for ML. The overall goal is for ML to be an enabler of sustainability transformation, rather than another cause of environmental impact.



Feature engineering and dimension reduction

Complexity: Medium - High | Impact: Medium | Scope: 10% of ML use cases

Feature engineering is the process of transforming raw data into features that are more informative and useful for machine learning models. This often involves domain knowledge and expertise. **One common type of feature engineering is naming embeddings, which are a way to reduce the dimensions of a dataset.**

Naming embeddings are specialized deep neural networks that take and output the same number of dimensions. Input data with many dimensions goes into the model, and the same number of dimensions go out, but with a reduced number of nodes in between layers. This forces the model to learn how to represent the same data in fewer dimensions and restore it from that.

The quality and the size of your data are the most important factors in determining CO_2 emissions from machine learning. By applying proper feature engineering, data scientists can reduce the number of input dimensions to a model and therefore the overall size of the model. Smaller models require fewer to train and run, which **can have a positive impact on sustainability**.

Experiment tracking

Complexity: Easy | Impact: Medium - High | Scope: 90% of ML use cases

Building and training a machine learning (ML) model can be a highly iterative process. **To produce a well-performing model, data scientists need to test, change and optimize many aspects, including data preparation, model architecture, optimizers, learning rates, dropouts, other hyperparameters**. This requires an iterative process to test and optimize this.

In these iterative cycles, it is important to track experiments and their results. This can prevent data scientists from having to re-run expensive training jobs simply because they forgot the results of a previous execution.

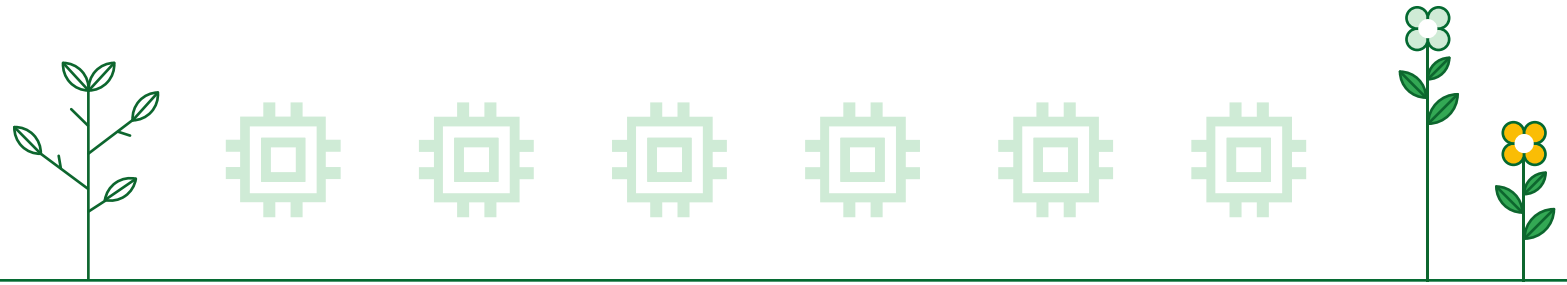
Vertex AI's experiment tracking feature can help data scientists in this task[1]. Additionally, there are tools, like Vizier[2] that can reduce the number of experiment cycles required hyperparameters by intelligently exploring the hyperparameter space.

Leverage accelerators

Complexity: Medium | Impact: Medium - High | Scope: 70% of ML use cases

Most data scientists already use accelerators in the cloud to speed up their training, **which typically leads to carbon reduction**. Yes, GPUs consume more power than CPUs, but they can train models much faster, which reduces the overall resource consumption of CPUs, GPUs, memory, storage and other resources.

Leveraging accelerators in the cloud gives you the flexibility to choose the best region and time to run your training jobs, so you can maximize the use of renewable energy. Google Cloud offers a specialized ML accelerator called the Tensor Processing Unit (TPU), which can further speed up training iterations. TPUs are available in many Google Cloud regions, including Iowa where they run at over 90% carbon free energy[3].



Post training quantization

Complexity: Medium | Impact: Medium - High | Scope: 50% of ML use cases

Once a model is trained, it is typically moved into production, where it is served for inference. Depending on the use case, a model may be called a very high number of times. For example, a visual inspection model in a manufacturing environment may be called thousands of times per minute to identify errors on production lines. Or a credit card fraud detection model may be called to inspect every transaction.

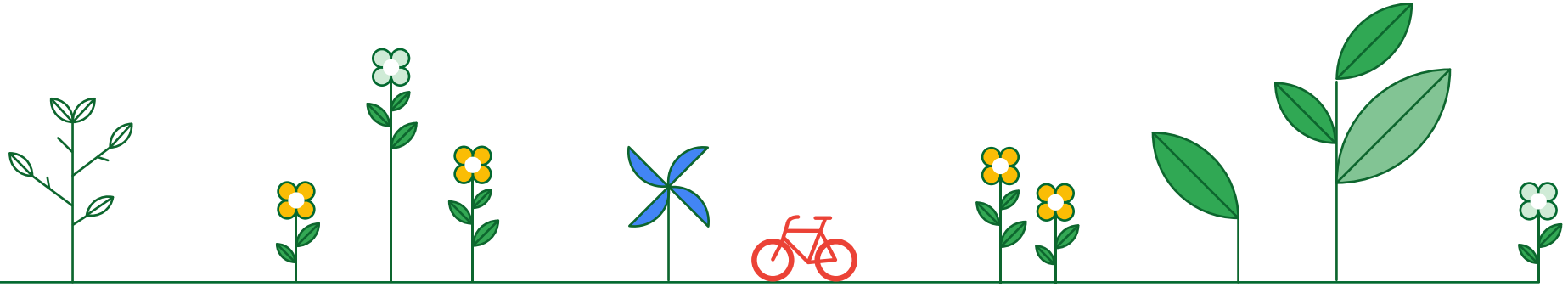
In such use cases, the power consumption of the model during inference can be significant. This should be considered from the beginning and may impact the model architecture, leading to a simpler and more efficient design.

Another way to reduce resource consumption during inference is through post-training quantization. This process converts the model weights into types (e.g. int8) with a reduced bit size. This reduces the size of the model, increases its speed, and reduces resource consumption, with a minimal reduction in precision[4].

At scale, when the model is inferred often, this can significantly reduce the carbon footprint of the model during use time.

The increase in organizations setting sustainability goals to achieve net-zero has increased during the last few years. Customers reach out to us to explore ways to measure and reduce the emissions associated with their IT infrastructure. Google Cloud has been the leader in providing reporting tools, as well as methodologies to make a more efficient use of their compute resources. This will lead to a reduction in their emissions that will help them achieve their sustainability targets.

Ester Morales
Customer Engineer, Sustainability Specialist



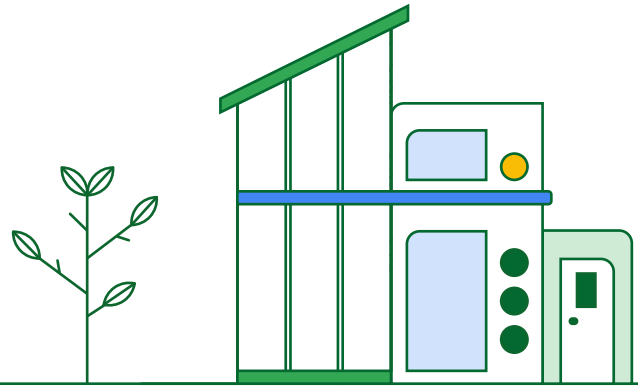
Footnotes

1 Track, compare, manage experiments with Vertex AI Experiments, blog, Google Cloud, 13 July 2022

2 Vertex AI Vizier overview, Vertex AI guide, Google Cloud

3 Cloud TPU v4 records fastest training times on five MLPerf 2.0 benchmarks,
AI and Machine Learning blog, Google Cloud, 30 June 2022

4 Post-training quantization, Model Optimization Guide, Tensor Flow



cloud.google.com

Go Green Software

A guide for Data Engineers

Google Cloud

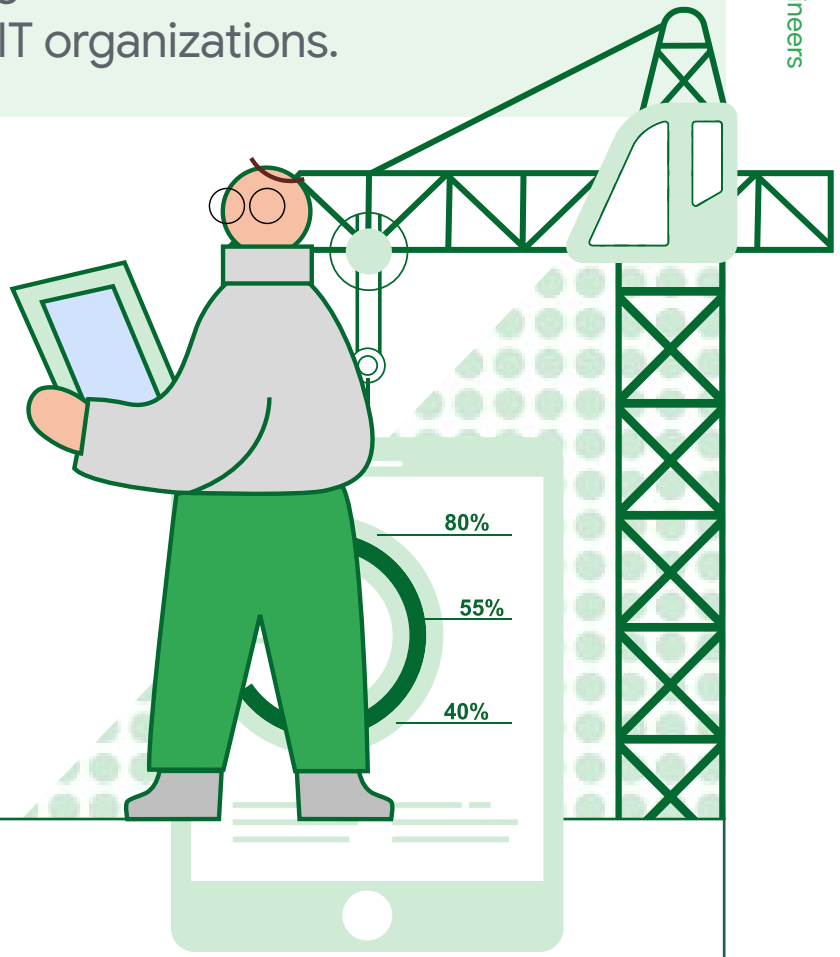
Contents

Introduction

Data plays a key role in today's business processes and systems. Sourcing, ingestion, processing, preparation and using increasing amounts of data has become a crucial part of the data engineer role in IT organizations.

Data engineers serve the rest of the business with the perfectly prepared data they need, to make informed decisions. Data engineers have a deep understanding of data, structures and statistics, and they know how to use the right tools to make data useful.

As more data is processed, this can become a significant source of CO_2 emissions. Large pipelines are regularly spun up to process terabytes of data, and large amounts of data need to be stored for further analysis. All of this can be improved by following some practices described below.



Data lifecycle

Complexity: Easy | Impact: Medium | Scope: 40% of data

The usage pattern of data typically changes over time. In the creation phase, the data is often used often for analysis, reports or advanced ML purposes. **As the data gets older, it is less utilized**.

For example, financial data is heavily used for reporting during the financial year it was created and until the FY investors report is out. After that it is still used occasionally for year-over-year comparisons, but after two or three years it is barely used, but needs to be stored for at least 10 years for regulatory reasons.

Considering this lifecycle and the usage pattern, **data can be moved to specific long-term storage**. Cloud has options for cheaper long-term storages, for example coldline or archive storage class. Due to the specialized nature of these services, they require less resources and energy and therefore have lower CO_2 emissions.

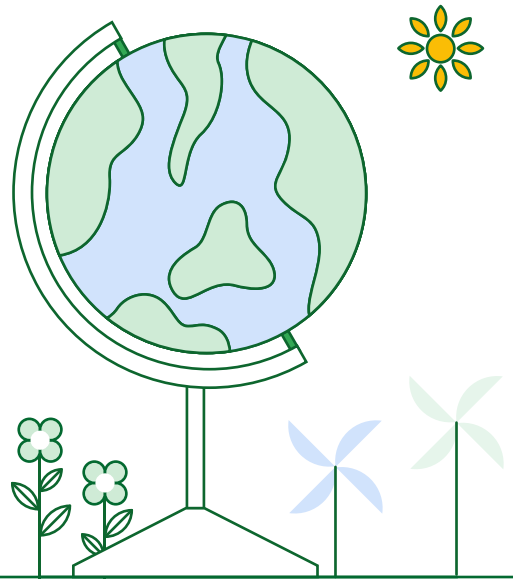
More complex data lifecycle management approaches are driven by the aspect of raw vs. aggregated data. In many cases, it is sufficient to keep only an aggregated version of the data for common usage, while moving the raw data to a more long term storage approach with a reduced footprint. Pre-aggregating the data would not only allow this reduction, but would also reduce the compute power required to aggregate the data ad hoc.

Data locality

Complexity: Easy | Impact: Medium | Scope: 60% of data

The Carbon Free Energy (CFE) % and the carbon intensity of the electricity grid vary by country. **Data storage and access also require electricity, so the location of data is an important factor in CO_2 emissions**.

If data can be moved, this should be considered. For example, Google Cloud provides all the details about CFE% and carbon intensity to help you make decisions[1]. It should also be possible to move the processes that use the data. However, if moving data to another location increases network traffic because the data creators and users are located elsewhere, this traffic also needs to be evaluated. Estimating network traffic is challenging, but projects like The Shift Project has estimated an average number of kWh per GB[2].

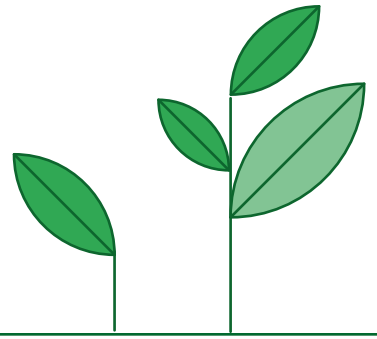


Batch job optimization

Complexity: Easy - Medium | Impact: Medium | Scope: 20 - 30% of data

Batch jobs are often used to process large amount of data on a regular schedule. However, these jobs can be inefficient if they are executed too often, at the wrong times, or without considering the amount of data being processed. There are a few aspects that can be improved with regards to classic batch jobs, regardless of the technology that is being used.

Frequency: Review the schedules of all batch jobs to determine if they need to run as often as they currently are. With a complete overview of all batch jobs and their dependencies, you can identify optimized timing, with fewer executions. **It is also important to gather feedback from the end users to understand their data freshness requirements.**



Timing: The carbon intensity of the electricity grid changes over time and by region. **Batch jobs should be executed at times and places where the lowest carbon intensity is available**. Vertflow[3] is a useful tool for automatically scheduling Cloud Run executions as part of an Airflow pipeline based on the carbon intensity.

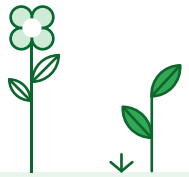
Amount: Instead of processing all of the data with every run, consider a delta approach that **only processes the data that has changed since the last execution**. This requires more upfront design and implementation work, but it can significantly reduce the required resources.

These are the basic practices to consider when optimizing batch processes. It is important to measure your current resource usage and track the impact of your organizations.

One potential alternative to batch jobs is to use BigQuery's materialized views. **Materialized views pre-calculate and store data for faster access, and they use a variety of optimizations to reduce overall resource usage**. For example, when the base table is updated, the subsequent data points are automatically updated incrementally rather than fully. This makes materialized views a highly effective alternative to batch jobs for many tasks that can be solved with SQL.

Event-driven

Complexity: High | Impact: Medium | Scope: 20% of data



The next level of batch job optimization is an event-driven approach. **Instead of running scheduled data processing jobs, the new data can be processed immediately as it arrives**. Messaging queues and serverless compute can help realize this pattern.

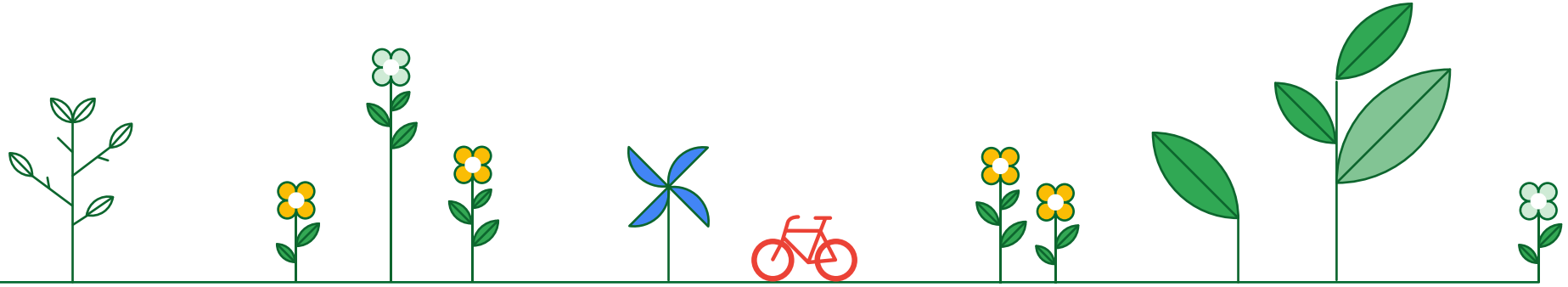
However, transforming batch jobs to an event-driven architecture can be a significant undertaking. It is important to understand the source systems that generate the data and how they **can be changed to push data, rather than requiring another tool to pull it**. Change Data Capture (CDC) mechanisms are a good starting point for identifying the capabilities of source databases[4].

Typically, the data processing and transformation steps also need to be changed. Apache Beam is a framework that converges batch and stream processing, providing a flexible pipeline approach for both. There are other tools available as well.

Between the data creators and users, there is typically a publisher and subscriber mechanism, implemented using a message queue. This can create larger publish and subscribe cycles, but it also provides clear lineage of data and its usage.

Many Google Cloud customers and partners are involved in green software generally and sustainability is a key topic for business of all sizes. That said, education is essential for the growth of green software in the future. This is why Google is committed to sharing information with IT professionals about the carbon footprint of our development environments and services. By making this information available, we hope to help engineers make more informed decisions about their work and to create software that is more environmentally friendly.

Charlotte Hutchinson
Sustainability Specialist, Google Cloud



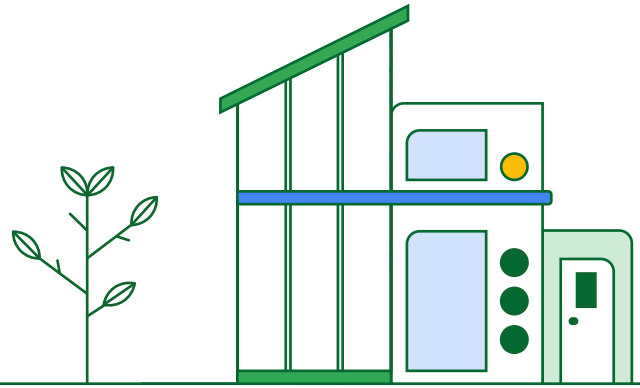
Footnotes

1 Carbon free energy for Google Cloud

2 The Shift Project

3 VertFlow 0.4.1

4 Change Data Capture(CDC) in PostgreSQL, Ramnesh Naik, 18 April 2018, Medium



cloud.google.com

Go Green Software

A guide for Data Architects

Google Cloud

Contents

Introduction

Data Architect should consider **the internal landscape of complex data platforms and enable data transformation strategies across the organization.**

They are responsible for accelerating, de-risking, and expanding the data transformation journey, pushing organizations to review the required data capabilities and define the data architecture. They provide an **opinionated view on cross-cutting topics** such as data governance, data mesh, understanding and adopting data technology stacks and complex landscapes, advising on design patterns and best practices. With respect to the cloud, they should also understand cloud economics, sustainability and the emissions impact of their data cloud environment.

In the 2010s, many organizations were embracing the ‘Era of Big Data.’ This was in response to the ease with which organizations could collect, process and store vast amounts of data generated from their enterprise systems, products and customers. This was largely facilitated by the shift to cloud, where storage was readily accessible and relatively cheap.

The data estates of organizations grew significantly, often under the mantra that ‘more data is better’. As big data platforms demand large storage and computational resource requirements, data architects have significant responsibility in the design of sustainable data solutions to reduce energy demands and create efficient strategies for data processing and consumption.

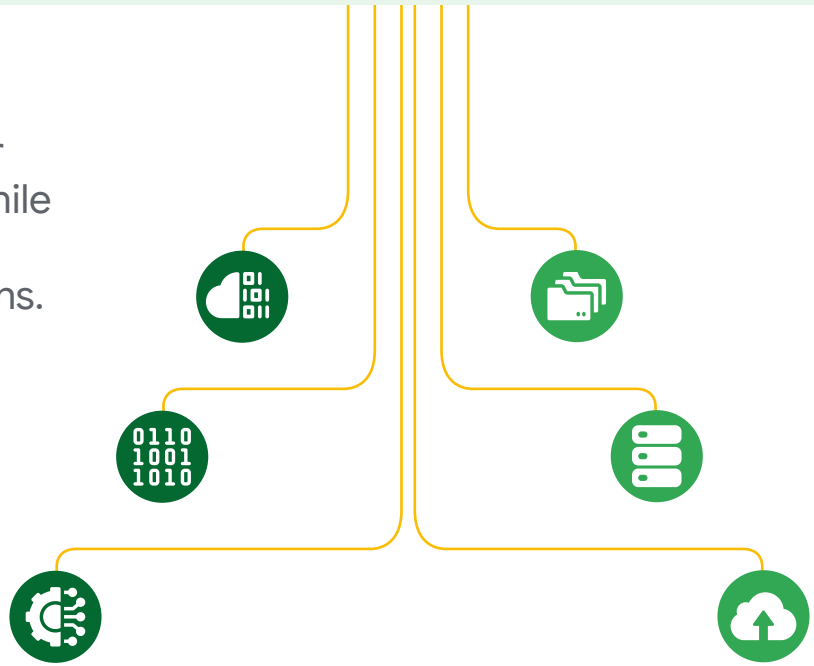
Decoupling storage and computation

Complexity: Medium | Impact: Medium-High | Scope: 30-40 % of data and computation resources

Monolithic data architectures typically have strong hardware and software dependencies between storage and computation.

This is a common feature in on-premise environments, where clusters of servers have persistent disks (or other storage solutions) directly attached to the machines. While this improves performance, it introduces a dependency that can be addressed by adopting cloud-native solutions.

To break this dependency in the cloud, we recommend starting with an assessment of storage and compute resources. Perform a series of technical interviews with the data architecture department to identify the target solution that best fits for each specific case. This approach helps to **dramatically reduce migration and optimization risks.**



Step by step guide: Storage

1. List the type of data stored in a data platform (structured, unstructured, semi-structured).
2. Define compliance requirements(e.g. Personal Identifiable Information (PII), anonymous, sensitive information).
3. Identify data security mechanisms (e.g. how data is encrypted, who manages encryption keys, rotation rules).
4. Map requirements with right storage products (e.g object storage, relational db, NoSQL db, time-series db, data warehouse storage).

Step by step guide: Computation

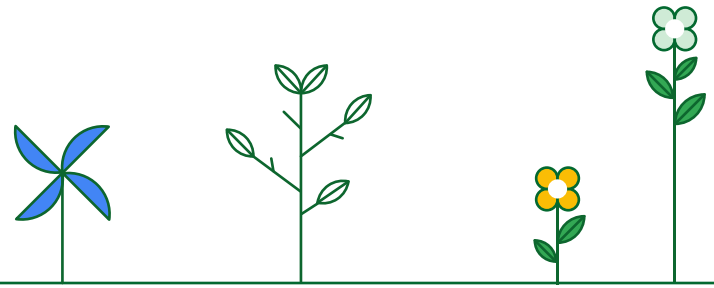
1. Collect information about cluster's topology (e.g. number of VMs/server, node specialization, CPU, Ram).
2. Collect benchmarks (e.g. peak of requests by user space, by time, by data volumes processed).
3. Map capabilities to compute products, looking to modernize infrastructure as much as possible (containers and serverless technologies for data).

Data governance

Complexity: Medium-High | Impact: Medium-High | Scope: 20-40 % of data

With big data platforms, manual data enhancement and policy execution becomes nearly impossible, inevitably leading to policy violations and turning data platforms into a data swamp. **Data governance methodologies based on cloud technology allows organizations to automate data management and enable governance at scale.**

Building an enterprise data catalog and defining a data management strategy play a key role in understanding what data is owned by who within the company. Users who know the organization's data, what it means, and how it should be used enable companies to shape their strategy based on data-driven culture. The introduction of concepts like data ownership, recoverability, lineage, and data quality policy makes data platforms trustworthy. This means that users become confident that they can use the data to meet their needs.



Typically, companies choose between two main models of data governance:

- Centralized approach
- Federated approach

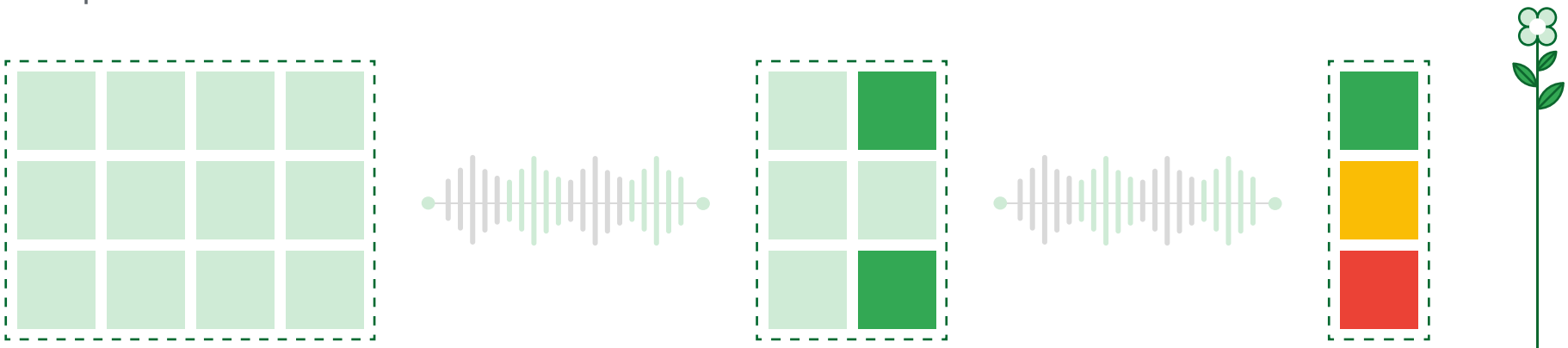
Data governance strategies and policies can have a significant impact on how data is processed and stored and the relevant emissions associated with that architecture. For example, federated approaches will avoid data replication and transfer and limit the carbon footprint of that data for the organization.

Single source of truth for data

Complexity: Medium | Impact: High | Scope: 40% of data

A data management strategy is essential for large data platforms to keep data discoverable, democratically accessible and of high quality. This involves **identifying the single source of truth for each data set**, avoiding duplicates, and resolving ambiguities.

This phase has a significant impact on data storage volumes and helps to implement data life cycle management. Reducing storage volume means reducing the storage carbon footprint of data platforms.



Step by step guide

1. Cleanup datasets means keeping track of all the copies of the same dataset created for different teams or stakeholders, tagging them with the purpose to eliminate later those that are useful.

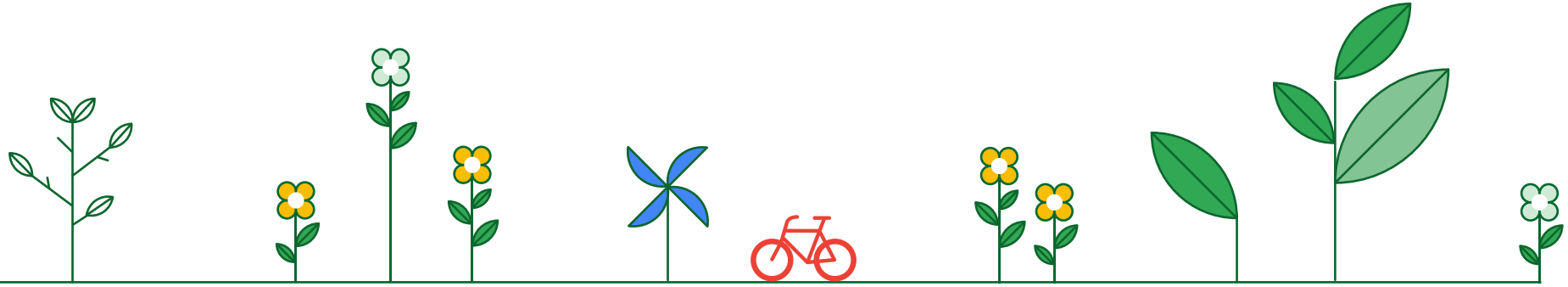
2. During data cleansing, keep the data catalog updated with all the business and technical metadata related to the datasets (e.g. data owner, update data, access policy, column type, data description, versioning).

3. Keep stored only the ones that have a clear lineage reported within the data catalog and delete all remaining copies of data.

4. Introduce a data lifecycle management to keep data platform cleaned.

By 2030, Google aims to achieve net-zero emissions across our entire operations and value chain. This is supported by an ambitious clean energy goal to run on 24/7 carbon-free energy on every grid where we operate. We see green software as a critical component to help us achieve these ambitious goals. Energy efficient software has always been a priority for Google, and with our leadership in AI, we have focused on driving an optimized stack for growing compute demands. We are excited to share our best practices through the green software principles so that we can help our partners reduce the carbon footprint of their own software solutions.

Savannah Goodman
Data and Software Climate Solutions Lead, Google Cloud



cloud.google.com