

the
GORILLA
GUIDE[®] to...



Navigating Churning Data, Analytics, and AI/ML

How to Deal with Data Overload

DAN SULLIVAN, PH.D.

Navigating Churning Data, Analytics, and AI/ML

Dan Sullivan, Ph.D.

TABLE OF CONTENTS

Introduction.....	1
AI and Analytics: From Data to Intelligence.....	1
The Need for Data Engineering.....	7
From Data to Models.....	11
The Lifecycle of AI and Analytics Models.....	14

Copyright © 2022 by ActualTech Media

All rights reserved. This book or any portion thereof may not be reproduced or used in any manner whatsoever without the express written permission of the publisher except for the use of brief quotations in a book review. Printed in the United States of America.

ACTUALTECH MEDIA

6650 Rivers Ave Ste 105 #22489 | North Charleston, SC 29406-4829

www.actualtechmedia.com

Publisher's Acknowledgements

EDITORIAL DIRECTOR

Keith Ward

DIRECTOR OF CONTENT DELIVERY

Wendy Hernandez

CREATIVE DIRECTOR

Olivia Thomson

SENIOR DIRECTOR OF CONTENT

Katie Mohr

PARTNER AND VP OF CONTENT

James Green

ABOUT THE AUTHOR

Dan Sullivan, Ph.D., is a principal engineer and architect focused on cloud architecture, data science, machine learning, and data architecture. He's the author of six books, most recently on cloud and NoSQL databases, as well as several online courses on machine learning, data science, and cloud computing that have had over 1 million views. He's the author of Google Cloud's official certification study guides out from Sybex. Follow Dan on LinkedIn at www.linkedin.com/in/dansullivanpdx/.

Introduction

Welcome to The Gorilla Guide® To.. Navigating Churning Data, Analytics, and AI/ML, Foundation Edition. This book couldn't come at a more opportune time. Data is being created at a head-spinning rate, and will continue to explode into the future. What to do with all of it?

Make it work for your organization! That involves leveraging artificial intelligence (AI) and machine learning (ML), which are enabling the massive scaling of services that were, until recently, performed by humans. Analytics is helping decision makers better understand their customers, organizations, and market trends.

An essential component of both AI and analytics is their dependency on data, including data not normally collected for business operations. For example, businesses may now have access to geolocation data collected from mobile devices, web browsing history, and social media.

AI and Analytics: From Data to Intelligence

The amount of data being collected now is much more voluminous than what was routinely collected in the past, and frequently includes detailed personal data beyond simple financial, inventory, and customer order information.

With all this available data, the primary challenge is generating useful insights from it, and because the data is so diverse, no single discipline or technology can meet every need of modern analytics.



According to [TechJury](#), the amount of data being created is staggering:

- 1.7MB of data is created every second by every person during 2020
- In the last two years alone, the astonishing 90% of the world's data has been created
- 2.5 quintillion bytes of data are produced by humans every day
- 463 exabytes of data will be generated each day by humans as of 2025
- 95 million photos and videos are shared every day on Instagram
- By the end of 2020, 44 zettabytes will make up the entire digital universe
- Every day, 306.4 billion emails are sent, and 500 million tweets are made

ANALYTIC INSIGHTS FROM DATA

Analytics is a term that describes the application of several disciplines, techniques, and tools, including databases and query languages, statistics, reporting and visualization, and anomaly detection to reveal meaningful patterns in data.

Databases and Query Languages

Let's first look at databases and query languages. Relational databases are widely used within data analytics. SQL, the standard language for relational database management systems, is particularly well suited to both ad hoc querying and standard reporting.

Because of this, many existing businesses applications use relational databases. In some cases, however, relational databases are not the best option. NoSQL databases are defined for specialized cases, and include key-value databases; document databases for semi-structured data; wide-column databases for large-volume data applications (especially those requiring low latency writes); and network databases, which are well suited to network analysis such as social connections and chemical interactions.

Statistics

The discipline of statistics has two primary divisions, descriptive and inferential, both of which are helpful for understanding large data sets. Descriptive statistics is useful for describing overall patterns or properties found within data, including statistical measures like mean, median, variance, and standard deviation. Inferential statistics lets you use subsets or samples to compute descriptive statistical information.

For example, the average of a sample is likely a good approximation of the average of the full population. This kind of statistics is also useful for hypothesis testing, such as the question, "Is one subset of customers more likely to make

a high-valued purchase in the next 60 days than another subset?” Inferential statistics helps measure the likelihood a given hypothesis to query is true.

Anomaly detection is used to detect outliers, such as unexpected changes that are likely fraudulent, a sudden surge in resource utilization on a network, or an unusual drop in sales in a particular store.

Reporting and visualization

Reporting can be quite beneficial for management in supporting its decisions. With reporting tools, managers can gather reports on sales, revenues, and profits—information that’s indispensable when making decisions about business strategy and tactics.

Additionally, business analysts can explore data, look for patterns, and gather insights with ad hoc query tools and data visualization.

Anomaly detection

Finally, anomaly detection is used to detect outliers, such as unexpected changes that are likely fraudulent, a sudden surge in resource utilization on a network, or an unusual drop in sales in a particular store.

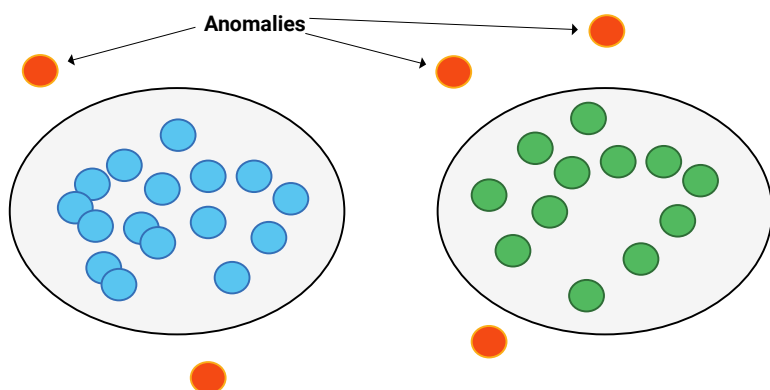


Figure 1: Example cluster anomaly

For example, if you use cluster techniques to group data, data that doesn't clearly fit into established groups may be an anomaly worth investigating (see **Figure 1**). When working with time series data, you can predict potential data values at future times, and actual data that's far off from the prediction may be worth examining.

The tools of analytics are designed to help us better understand what data is telling us. AI is a complementary discipline that focuses on developing techniques for solving specific kinds of problems using data.

INTELLIGENT INSIGHTS FROM DATA

AI is a broad field centered on developing software that can perform tasks that in the past have required difficult-to-encode human knowledge. ML is a sub-field of AI dedicated to developing methods that allow software to perform intelligent operations without requiring explicit human coding;

instead, software learns from data. These techniques are applied to a variety of practices:

- Classification is the process of assigning class labels or categories to items in data sets so they can be organized into groups. For example, classification allows a computer to distinguish a legitimate credit card transaction from a fraudulent one.
- Natural language processing utilizes ML techniques to allow software to understand and generate natural human language. It's a key enabling technology underpinning applications such as translators, document classifiers, chatbots, and virtual assistants.
- Machine vision teaches computers how to process visual information by training them to recognize patterns in images, such as distinguishing a bus from a car or identifying a quality control issue on a production line. Some of the more common applications of machine vision include facial recognition, object detection, and image classification.
- Segmentation is the process of grouping customers based on their similarities while clustering is the process of identifying similarities in customers in order to group them.
- Sentiment detection is the process of identifying tone and other subjective information based on text analysis and natural language processing. This is especially useful for analyzing reviews, customer comments, and related texts.

- Recommendation making is the process of suggesting information, products, or services a user or customer may be interested in. Recommendations can be based on data related to a subject's preferences, online activity, or the behavior of similar users.

The combination of analytic and ML techniques are driving innovations and insights but to keep those benefits coming, access to the right kinds of data must be provided. The practice of providing data is becoming more formalized and is commonly referred to as data engineering.

The Need for Data Engineering



The first step in gathering and preparing data for analytics and ML is identifying data sources. Business data comes in many forms and few are structured in ways to support analytics and ML.

Moreover, systems that create the data we analyze are often not well suited to supporting analysis. For example, an order processing system will be optimized for executing fast, reliable transactions, but the design choices that lead to an efficient transaction processing system don't lend themselves to analyzing large volumes of data.

Another factor driving the need for data engineering is that data is created and stored in many different systems and must be consolidated before analysis or ML can begin.

Another factor driving the need for data engineering is that data is created and stored in many different systems and must be consolidated before analysis or ML can begin.

Data engineering encompasses three broad types of operations: ingestion, storage, and transformation services, as shown in **Figure 2**.

DATA INGESTION

Extracting data from source systems and loading it into consolidated data stores can range from relatively simple operations on small amounts of data to logistically challenging operations moving large volumes of data.

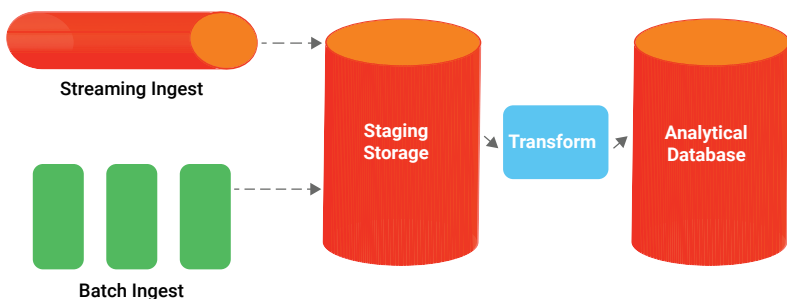


Figure 2: Ingestion, storage, and transformation processes

One common approach is to use batch jobs to load data into analytical databases. At regular intervals, data is extracted from source systems and loaded into the target database. It's possible in some cases to directly load data from the source to target systems, but it's common to stage the extracted data on file system or cloud object storage systems. This decouples the source and target and allows for more flexibility when designing the transformation processes.

When data needs to be analyzed as quickly as possible, streaming ingestion is a better option than batch loading. For example, sensors may send measurement data to an ingestion point where it's immediately processed and inserted into the analytical database.

When data needs to be analyzed as quickly as possible, streaming ingestion is a better option than batch loading.

Again, there are advantages to decoupling source and target systems, so data may be written to a queue or message-passing system that's better able to tolerate spikes in load than a database trying to keep up with high volumes of data.

When migrating large volumes of data, such as from an on-premises data warehouse to a cloud solution, batch uploads over a network may not be a viable option. Transferring 10PB of data over a 10Gbps network would take about 120 days. Cloud vendors provide transfer storage systems that can

be shipped to a customer's site and attached to their network where data is copied from the source system to the transfer devices. The device is then shipped back to the cloud vendor and the data is copied from the device to a storage service in the cloud.

STORAGE SYSTEM DESIGN

Deciding how to store analytic and ML data sets is a multifaceted challenge. Object storage is a good option for staging data when operating on entire files, while block storage and file system storage is needed for operating on data within files.

There's also the question of what type of database to use. Relational databases work well for some data science use cases, particularly when dealing with complex data models. NoSQL databases can be a better option when working with less-structured data. Analytical databases designed for highly scalable distributed query processing are quickly becoming the workhorses for large-scale analytics projects.

Deciding how to store analytic and ML data sets is a multifaceted challenge. Object storage is a good option for staging data when operating on entire files, while block storage and file system storage is needed for operating on data within files.

TRANSFORMATION SERVICES

In addition to moving and storing data, we also need to transform it into a form suitable for analysis. Prior to implementing transformation logic, it's a good practice to validate data. This can include checks to verify the amount of data in a data set as well as data-quality tests to identify missing or invalid attributes.

Transforming analytical data sets can be compute-intensive, especially when joining large data sets or repeatedly querying large tables. Saving intermediate results as materialized views or temporary tables can often eliminate some costly, repetitive querying.

As well as designing ingestion, storage, and transformation systems that meet functional business requirements, data engineers also design for scalability, reliability, and durability.

Many of these data engineering practices used in analytics are also used in ML and other AI operations.

From Data to Models



The democratization of ML and analytics is increasing. More tools are available to support quantitative specialists like data scientists and statisticians, but also to help business analysts and other domain experts. This is creating an opportunity to apply AI and analytics to support less quantitatively oriented users.

Expect to see:

- More tools to support data specialists, especially around the deployment and monitoring of data science and ML models.
- As data sets grow, so does the need for computing resources. Specialists will continue to depend on tools and data management services provided by public clouds.
- Less need to build ML models from scratch. For example, pretrained models for vision and natural language processing will provide a starting point for many applications.
- The process of identifying useful data and choosing appropriate ML algorithms will be increasingly automated. This is perhaps the single largest driver to the democratization of ML.

As with any predictions, it's important to note assumptions and caveats. Even with pretrained models and automated ML, there's a need for data engineering expertise and an understanding of necessary infrastructure.



Management consulting firm Bain & Company wrote a [brief report](#) outlining six unintended consequences of AI and how to handle them. One of the report's assertions is that people actually become more important, rather than less, in this new reality. "In a world shaped by AI, human leadership matters more than ever," it concludes.

While automation will help increase the reach of AI, it will also increase the range of any unintended consequences of AI. Some of the most challenging aspects of ML require deep domain expertise, especially assessing fairness and detecting bias in models.

The democratization of ML and analytics is further supported by the adoption of sound data management practices, including data catalogs and data governance regimes. Data catalogs allow users of data to understand what kinds of data are available, the level of quality in that data, and how data sets relate to one another.

Effective governance of data and algorithmic decision making is crucial for the long-term effectiveness of AI and analytics. The challenges of training models to make fair and accurate predictions are well known within the ML community, but issues of fairness are also the topic of discussion in non-specialized public forums. Algorithms that make decisions about lending or healthcare will need to meet the same level of fairness and competence we'd expect from human practitioners.

The Lifecycle of AI and Analytics Models

AI and analytics models are introducing new practices in the way we use and manage information technology. It's no surprise, then, that some practices that have served well in the past may no longer be needed or should be adapted to emerging data engineering needs.

LONGER DATA LIFECYCLES

For example, it's been a common practice to purge old data to reduce storage costs and management overhead. With the advent of ML, this doesn't always make sense, since we don't always know what data will be needed in the future.

Consider an ML model that's used to find defective parts produced by an injection molding process. The model may work well with most objects, but not with parts that have certain characteristics. One way to address this is to train the model on more examples of parts with those problematic characteristics—one place to find those training examples may be in old data sets. This creates the need to support expanding data stores.

FINDING NEEDLES IN DATA HAYSTACKS

Maintaining large data stores can prove useful when building ML data sets, but the utility of the data depends on how readily data and ML engineers can access it. The metadata about

data sets needs to be tracked to help users identify data sets that may be useful to them.

In addition, it will be important to provide data-quality measurements about the data sets so users can make informed choices about which data sets to use and how much time to plan for transforming problematic ones. Users will also need sufficient information in order to understand how to integrate data from disparate data sets.

INCREASING IMPORTANCE OF UNSTRUCTURED DATA

Analytics is no longer limited to working with highly structured data like that found in relational databases. Business data now includes documents, images, videos, and audio. Messages sent to customer support groups are now being “read” by AI models that can also respond to the customer directly, without human intervention.

Similarly, ML models developed for audio can be the foundation of automated voice response systems. Automated video monitoring of production processes can help identify problems in operations without having to wait for a human to learn about the problem. Unstructured data complements structured data, and the two will increasingly be used together.

PROTECTING DATA

The need to ensure the confidentiality, integrity, and availability of data is certainly not new, but it can become more challenging as data volumes grow and data is copied from one system to another. Differences in information security practices between these systems can leave data vulnerable to compromise.

There are well-established practices for protecting data that should be part of data lifecycle management, including:

- Encryption to preserve the confidentiality of data both at rest and in transit
- Access controls to provide access to data only to users and services with a specific role-related need for that data
- Policies for defining rules governing access to resources
- Key management for creating, managing, and retiring encryption keys

As the scale of data grows, expect to see greater automation in the implementation and management of these practices.

With growing volumes of data, expect to see changes to storage systems to better meet the needs of analytics and ML users. Object storage will start to provide some of the functionality available in block and file system storage. Storage systems will improve on how they automate data lifecycle management tasks, such as migrating infrequently accessed data to lower-cost storage and optimizing how databases access data.

THINKING LIKE A DATA ANALYST

It's easy to forget the most important part of an analytics and ML practice is the human practitioners. Algorithms, models, and tools are essential to creating a data analytics and ML practice, but humans are the ones who devise the questions and formulate the hypotheses. Expect to see more demand from employees for training in data analytics, ML, statistics, and other forms of quantitative reasoning.

GET OFF THE FENCE

As you've seen throughout this Gorilla Guide, data is undergoing more than just a transformation—it's a revolution, led by AI and ML, which are in turn driving analytics that allow companies to glean insights that can lead to better products, happier customers, and a healthy bottom line.

If you're still on the fence about leveraging these new technologies, it's high time to hop off and start using them. You don't have to start big—slow and steady could be just the right thing for your needs. Then as you see the increased benefits rolling in, you can upscale your efforts and watch them pay off big for your organization.

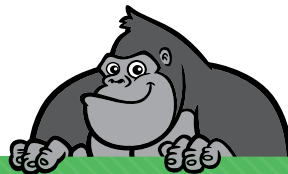
About ActualTech Media



ActualTech Media is a B2B tech marketing company that connects enterprise IT vendors with IT buyers through innovative lead generation programs and compelling custom content services.

ActualTech Media's team speaks to the enterprise IT audience because we've been the enterprise IT audience.

Our leadership team is stacked with former CIOs, IT managers, architects, subject matter experts and marketing professionals that help our clients spend less time explaining what their technology does and more time creating strategies that drive results.



If you're an IT marketer and you'd like your own custom Gorilla Guide® title for your company, please visit <https://www.gorilla.guide/custom-solutions/>