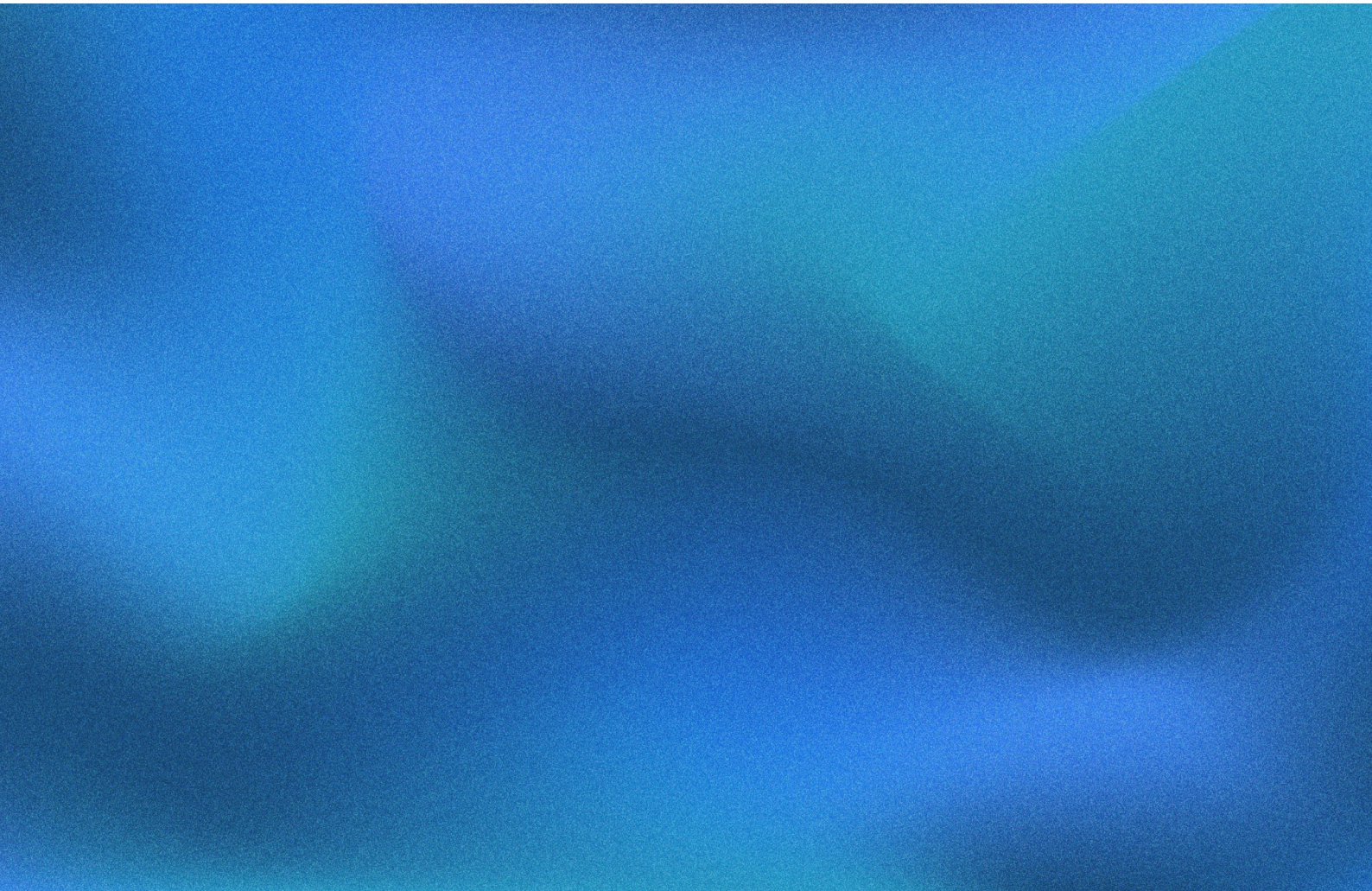
BRAIN POWERED AI COMPANY

AIDA takes the mess out of data strategy.

Faster. Easier.
Cost-effective data collection, processing, and access.



Content

01

8 common challenges that almost every company faces when handling data

From the business perspective, most companies struggle with connecting all the available data and extracting value from it efficiently.

From the technical perspective, handling data and preparing it for analytics and AI/ML experiments can cause serious bottlenecks for most companies if not planned for correctly.

We've categorized the challenges and here's the list of them:

1.1. How to connect all data quickly?

Collecting large amounts of unsorted data at high velocity from many sources requires many integrations. It means there are many **data sources** to assemble in **one place** to have an overview of business, which is time-consuming - oftentimes more than you can afford.

1.2. Where and how to store all that data?

We need to store all data to enable deep insights. Yet, not all data is ready for use - and systems store many databases in silos. This imposes a challenging task on the data storage layer. It needs to support **scalability** while ensuring c**ost efficiency** and **performant data access**. Dealing with a large amount of data daily puts pressure on data retrieval. Some need hot data (recent history) while others need cold data (past history).

1.3. How to organize data?

The stored data needs to be processed in order to be used for a range of tasks, such as data analytics and machine learning. That includes data cleansing, transforming, aggregating, feature extracting, and many more data verbs that induct structure to data.

All this **takes significant time**, and it's **prone to human error.** Result needs to provide efficient execution on more structured big data sets while staying flexible and fostering fast and innovation cycles.

1.4. Where should storing and processing happen?

The costs of maintaining physical networks rapidly become prohibitively expensive, so companies often cannot sustainably store or operate with big data sets on physical hardware.
Cloud-based solutions offer **scalable computing power**, which can keep up with the ever-increasing size of data volume. The problem is to strike a s**ubtle balance between cost-effectiveness and scalability of infrastructure**.

This requires a fine mix of testing, experience, and expertise.

1.5. How to keep up with a fast-growing business?

Business **is growing** - and with it, the **data, too**. Data infrastructure and teams need to **scale** together with business and data to ensure that they meet data insight needs.

Stakeholders need access to data to make decisions, derive insights, and create value. Company switches from 10 to 100 reports per day, or from once a day to once an hour, in the **blink of an eye**.

You need flexible solutions.

1.6. What about security and governance?

On top of everything, take **security** and **governance** (overall data strategy) into account.

When dealing with **sensitive information** in large systems, such as banks, security of user personal information or business-risky data is mandatory. Many organizations leave these decisions for later to maintain focus on infrastructure development. But data is like a child: it goes through countless transformations and it needs guidance and governance throughout that process.

If a company isn't involved in defining security and governance measures from the beginning, it is much harder to steer them later.

The problem (and solution) lies in streamlining this process.

1.7. How to get the value of the data?

When provided with unified storage, processing, and access to all data regardless of their nature, we can:

- Use data in **business intelligence**, visualization, and reporting tools, in order to **derive deep insights** into business.
- Data scientists and analysts can find hidden patterns and conclusions to **enable data-driven business.**
- As data grows, business can grow too by **monetizing collected data**. You can sell non-proprietary data, offer data as a service and much more to fuel the revenue streams.

1.8. How do we get more value?

When working with enormous amounts of data, **machine learning and AI** are the way to go in order to get even more insights, more hidden patterns, and automate many business processes.

Let's imagine you're trying to implement a recommendation engine without solving any of the challenges above. Everything becomes a manual process and turns into a lengthy preparation. You need **months** to create an environment for a useful recommendation engine - let alone implement it and see real value from it.

Have you recognized your organization in any of these questions?

Having listed all this, it becomes apparent that data teams need a **well-established and automated way** of creating, updating, and deploying ML and AI models in production. The entire data strategy relies on **heavy orchestration** of all these processes.

02

Introducing AIDA – an AI-ready Data Platform

AIDA is an AI-ready Data Platform solution that enables collecting heterogeneous data from various data sources, storing and processing it in a scalable and cost-effective way.

It enables data to be accessed in an easy and performant manner, enabling integration with standard BI tools and allowing data scientists to experiment and apply AI algorithms and models.

Since it's a c**ollection of ready-made functional blocks** whose combination supports many use cases, you can set it up and harness the value from the data much faster compared to available solutions on the market.

You can cherry pick the functional blocks needed for your use case, and such flexibility allows you to get valuable business insights based on the actual data **without disturbing existing operations**.

But it all takes time, and expertise, and tools, and integrations, and…

What if there's a way to reduce implementation time and get your data in an AI-ready state much faster - while gaining business value and saving the budget at the same time?

03 How AIDA solves these 8 challenges

"Get a strong and industry-agnostic data strategy tool"

Technically, AIDA brings together the most advantageous aspects of the two most commonly used data infrastructure systems - **Data Warehouse and Data Lake**.

- From a **Data Lake perspective**, it allows for collecting and working with both structured and unstructured data in various data formats, and relies on a scalable and cost effective data storage.
- From a **Data Warehouse perspective,** AIDA offers an unified and structured access to all data.

This provides you with a **Data Lakehouse concept**, on top of which AIDA offers features for applications of ML/AI, making it possible to extract additional value from the data through new insights and implementation of automated decisions.

- Decreased time to value. By the end of the first month of development, AIDA enables you to find insights from data. That makes your business process more efficient.

- Reduced costs. AIDA helps you reduce expenses by opting for serverless services rather than on-demand, selecting clustto the use-case and applying auto-scaling for unexpected load.

This design allows for adjustments to different sizes and for large amounts of data and processing to be handled without going over budget. It is a **cloud-agnosti solution**, possible to implement on-premise.

- **Central data repository that doesn't break the system**. When applied and customized, AIDA works as a central data repository. It allows you to collect, store, and process data in one place. Integration with data sources requires minimum modification of existing systems/applications.

- **Bulletproof compliance**. Metadata allows airtight **data governance.** AIDA is **GDPR compliant** and management of the roles for accessing the data can be easily done.

- **Full control of resources**. Its modular structure and approach make it possible to start with the essential components only and to build up gradually. That allows you to use this platform as you need.

- **Data access anytime**. It provides you with a **unified access point** - thus making the pre-processed and structured data easily accessible for both everyday business use, business intelligence, and AI experiments (*making it AI-ready*).

04 A bird's eye technical overview of AIDA

From a high-level perspective, AIDA comprises the following layers:

- ingestion,
- storage,
- processing, and
- access.

There are also two additional supporting layers - governance and infrastructure.

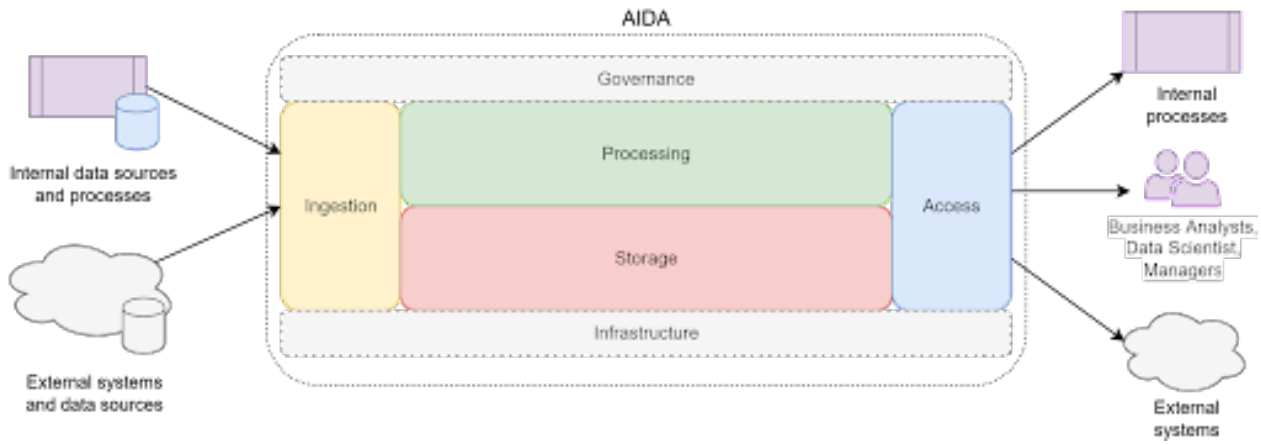


Figure 1. The conceptual diagram of AIDA layers

4.1. Ingestion

The Ingestion layer handles receiving large amounts of data with high velocity in various formats. This layer connects to different data sources and stores raw data in the Data Lake landing zone. Modern cloud data architectures use **AWS Glue, Azure Data Factory, GCP Dataproc** for ingestion.

4.2. Storage

The Storage layer uses scalable and cost-effective technologies to store large data amounts efficiently and provide performant data access. Modern data architectures rely on cloud object storage (**AWS S3, GCS, Azure Data Lake**) to implement this layer.

4.3. Processing

The Processing layer comprises batch and stream processing engines that handle data cleaning, transformations, aggregations, enrichment, feature extraction, AI model training, and real-time analytics.

It needs to provide flexible and performant execution on large datasets. Modern data architectures rely on open-sourced technologies, such as **Apache Spark** or **Apache Flink**, for batch and stream processing.

4.4. Access

The Access layer provides unified access to data in Data Lake, regardless of underlying formats and implementation details. It integrates with standard visualization, BI, and reporting tools.

Usually, the access layer supports SQL-compatible queries. Modern data architectures use **AWS Athena, Azure Synapse Serverless SQL,** or **Apache Presto/Trino** to provide SQL-like data access. Also, this layer takes care of the data access control.

4.5. Governance

Different teams should have different permissions related to data management.

The Data governance layer allows organizing and implementing policies and procedures to the data access control. It keeps the data platform secure and defines how the data is stored, gathered, processed and disposed of.

Data governance also implements external standards set by industry, governments, data consumers, and other stakeholders.

4.6. Infrastructure

Data infrastructure is the foundation of the modern data platform, which enables data storing, processing and exposing.

Automated infrastructure maximizes efficiency, productivity and makes collaboration much easier. Quick and flexible infrastructure setup ensures getting the business value out of the data faster.

We use Infrastructure as code (IaC) tools like **Terraform, Pulumi and AWS CDK** to implement this layer.

The Data Lake can store data in all stages of the refinement process. You could store raw data right alongside an organization's structured, tabular data sources (like database tables), as well as intermediate data tables generated during the refinement of raw data.

Unlike most databases and data warehouses, **Data Lakehouse can process all data types** — including unstructured and semi-structured data like images, video, audio, and documents — which are critical for today's machine learning and advanced analytics use cases.

4.7. Storage Layer

Delta Lake is an open format data management and governance layer that combines the best of both Data Lakes and Data Warehouses (Lakehouse architecture).

It solves the challenges of a Data Lake by adding a transactional storage layer on top.

A Lakehouse uses similar data structures and data management features as those in a data warehouse, but runs them directly on cloud Data Lakes. Ultimately, a Lakehouse allows traditional analytics, data science, and machine learning to **coexist in the same system, all in an open format.**

4.7.1. Key Delta Lake features

- **ACID transactions (provides serializability isolation level)**

Audit history and row-level updates/deletes help in possible regulatory requirements. ACID transactions ensure that every operation fully succeeds or aborts to maintain datasets in a consistent state.

- **Schema enforcement and evolution**

Schema enforcement & evolution solve frequent transactional schema update issues. Delta Lake provides DDL similar to standard RDBMS systems.

- **Data versioning (time travel)**

Data versioning is essential for ML experiments because it enables reproducible experiments. We can trace datasets used for ML models and debug model outputs if necessary. Also, we can time travel to previous data versions and learn from historical changes.

- Open data format (Parquet),
- BI tools operate directly on Delta Lake, thus avoiding multiple copies of data,
- Audit history,
- Support for row-level updates and deletes,
- Unified batch and streaming sources and sinks,
- 100% compatibility with Apache Spark,
- Performance: it uses optimization techniques such as data skipping, compaction, and highly accessed data caching.

Please note *- Delta Lake is not designed for online transaction processing (OLTP), even if it supports ACID transactions. Described data architectures, including Delta Lake, are suitable for analytics workloads (OLAP), separated from core transactional systems. However, Delta Lake can serve rarely accessed transactional data and reduce the load from transactional systems.*

4.7.2. 3 standard zones in the standard data architecture

A standard data architecture uses data zones that correspond to different quality levels in the data engineering pipeline, progressively adding structure to the data.

We define three zones:

- Bronze,
- Silver,
- Gold.

Bronze - A landing zone with raw data from multiple sources stored in native formats (CSV, Parquet, JSON, Avro, text, binary). There is no schema enforcement in this zone. Data can be structured, semi-structured, or unstructured.

Silver - Refined zone with cleaned, organised, and transformed data. Tables in the silver zone contain structured data with proper schema enforcement & evolution. Everything is SQL queryable and stored in Delta format.

Gold - Finest zone with business-level aggregations ready for serving. This zone contains enriched and well-organised tables with proper schema enforcement & evolution. Like the silver zone, everything in the gold zone is SQL queryable and stored in Delta format. Reports, dashboards and machine learning use data from the gold zone.

05 3 ways to implement AIDA

Depending on your infrastructure and business needs, there are three recommended ways to implement AIDA - by using:

- Databricks,
- AWS, and
- GCP.

The sections below reveal technical details and high-level overviews of infrastructure for each technology.

5.1. AIDA on Databricks (recommended)

The recommended approach to the AIDA implementation is to use the **Databricks** platform.

- Databricks platform offers managed tools for data transformation, governance, cataloging and querying.

- Resources used by the Databricks platform are deployed on the desired cloud account (AWS, GCP or Azure) and the data stays on the user's cloud account.

- Databricks is a unified platform on top of all three major cloud providers. It offers standardized ways of data management regardless of the cloud platform in the background.

This ensures better collaboration and productivity, and that's why this is the recommended AIDA implementation.

AIDA automatically sets up the data platform based on Databricks, which quickly becomes ready for bringing the value from the organization's data. Its modular nature enables choosing and provisioning only the subset of modules needed in the data platform to fit customers' needs.

Diagram below (Figure 2) shows the architecture of the Analytics platform based on the Databricks on AWS.

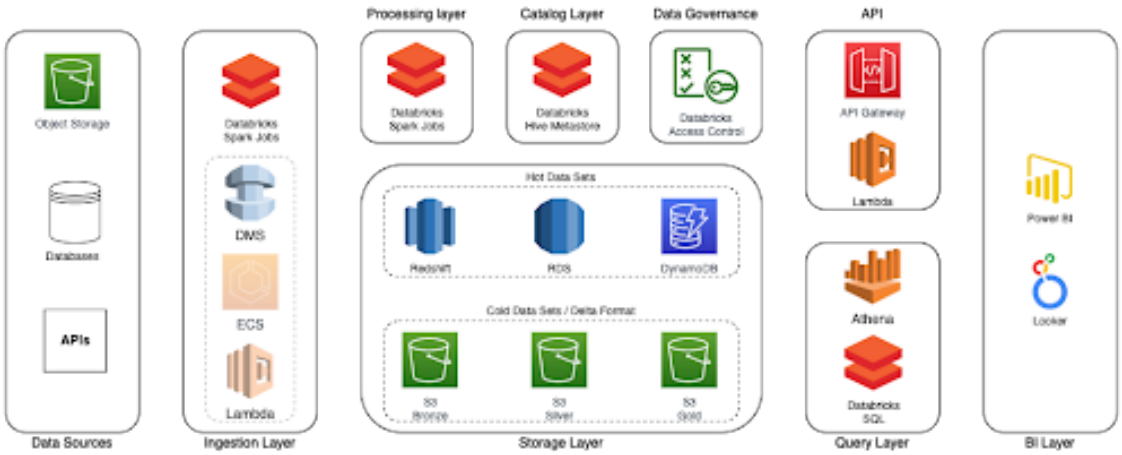


Figure 2. AIDA on Databricks

5.1.1. The Ingestion Layer

The purpose of the Ingestion Layer is to ensure reading the data from the data sources and writing it to the **Bronze zone** of the Data lake.

We can read sources in different ways. Some sources can be read using Databricks' managed Apache Spark*.

Note - Apache Spark is a distributed data processing framework that can quickly process an enormous amount of data.

We could also ingest some sources with AWS Native services like **ECS, Lambda, DMS**, etc. This completely depends on a source.

- **ECS** is a service that allows containers to run easily in the Cloud.

- **Lambda** is a serverless compute service which can run the code without the need to provision additional resources.

- **AWS Database Migration Service** (AWS DMS) is a cloud service that makes it easy to migrate relational databases, data warehouses, NoSQL databases, and other types of data stores. AWS DMS supports ongoing replication (CDC).

The Ingestion Layer reads raw data, and stores it in the Bronze zone of the data lake. The data format can be parquet or Delta.

5.1.2. The Processing Layer

The raw data from the Bronze zone needs to be processed and stored in the Silver zone in the appropriate format. The same applies to the Silver and Gold zones.

The Processing Layer handles data processing between the layers. We usually use **Apache Spark** as a processing framework. Depending on your business needs, we can process data in batches or in the real time. We can use Spark Structured Streaming processing to enable real-time analytics.

Similarly to the Ingestion Layer, we could use **Apache Spark, managed by Databricks**, in the Processing Layer. Instances used by Databricks are on the customer's AWS account.

5.1.3. The Catalog Layer

Catalog layer is a central repository of metadata. It contains metadata (schema and location) **for all data lake zones**.

This proves to be essential when we expose the data to the downstream systems, since it provides data consumers with access to the metadata.

Databricks has managed Hive Metastore (the most common metastore used in the industry) so it's almost an industry standard for a catalog layer.

5.1.4. The Governance Layer

Data governance layer requires implementation of the policies and procedures to the data access control. Databricks allow the implementation of the fine-grained access control policies for the data in all zones.

E.g. all your teams (e.g. Marketing, Management, Content, etc) could have a corresponding group, with appropriate permissions.

Delta Lake on Databricks can manage G**eneral Data Protection Regularization** (GDPR) and **California Consumer Privacy Act** (CCPA) compliance for the Lakehouse.

Since it requires deleting individual records, **Delta** is the right format to support these compliances. Delta optimizations can s**peed up point** deletions in the Lakehouse.

5.1.5. The Query Layer

Most data analysts and engineers are familiar with SQL, and most databases support SQL as a query language, so SQL is the most usual tool to query the data lake.

Databricks SQL is a managed service that offers a query layer for the Delta format. It is very performant and optimized for Delta lake.
Since SQL is used, data analysts will have the same interface as in any data warehouse. Databricks Serverless SQL offers instant computation and capacity optimizations that lowers overall cost.

5.1.6. The BI Layer

You could connect any popular BI tool to the Databricks Lakehouse. The most popular ones are Power BI and Looker.

- **Power BI** is an interactive data visualization software product developed by Microsoft, focusing primarily on business intelligence. It represents a collection of software services, apps, and connectors that work together to turn unrelated sources of data into coherent, visually immersive, and interactive insights.

- **Looker** is an enterprise platform for BI, data applications, and embedded analytics that helps you explore and share insights in real time.

5.1.7. API

Some data consumers cannot use SQL as a query layer for the platform. Another interface to the Lakehouse is the API.

APIs act as the "front door" for applications to access data, business logic, or functionality from your backend services.

AWS has popular and cost-effective serverless services that can serve data.

- **Amazon API Gateway** is a fully managed service that makes it easy for developers to create, publish, maintain, monitor, and secure APIs at any scale.

  Using API Gateway, you can create RESTful APIs and WebSocket APIs that enable real-time, two-way communication applications.

- **AWS Lambda** is a serverless, event-driven compute service that lets you run code for virtually any type of application or backend service without provisioning or managing servers.

5.2. AIDA on AWS – possible, same concepts, but different services

AIDA can also be implemented using the AWS native data engineering services. Data Lakehouse concepts from the previous approach are the same for this approach, but services that are used are different.

Diagram below (**Figure 3**) shows the architecture of the Analytics platform based on the AWS.

Figure 3. AIDA on AWS

5.2.1. The Ingestion Layer

Tools that are used for the ingestion layer in this approach are the same as in the previous approach. We can employ Apache Spark for some sources and AWS ECS, Lambda, and DMS for the other data sources.

The only difference between these approaches is that, with native AWS implementation, we would use the **AWS EMR Service**. AWS EMR service manages Spark cluster, so we use it for running Spark ingestion jobs.

5.2.2. The Processing Layer

In the case of the AWS Processing layer, we use Apache Spark for processing. AWS EMR manages Spark clusters. Additionally, in the case of architecture change and moving from the Databricks AIDA to the AWS AIDA, it is easy to migrate Spark jobs from Databricks to EMR, and vice versa.

5.2.3. The Catalog Layer

The Catalog Layer in the AWS AIDA is the AWS Glue Data Catalog. It is a native solution by AWS, compatible with Hive Metastore.

5.2.4. The Governance Layer

AWS does not have its own solution for data governance, so we implement the access control using AWS IAM roles and permissions. (!) This approach requires more time and effort to configure and enable data governance compared to the Databricks approach.

5.2.5. The Query Layer

AWS AIDA also provides the query layer with SQL as an interface.

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. Athena is serverless, so there is no infrastructure to manage, and you pay only for the queries that you run.

Athena natively supports the AWS Glue Data Catalog.

The AWS Glue Data Catalog is a data catalog built on top of other datasets and data sources, such as Amazon S3, Amazon Redshift, and Amazon DynamoDB. You can also connect Athena to other data sources by using a variety of connectors.

5.2.6. The BI Layer

As in the AIDA on Databricks, you can connect any of the popular BI tools in the industry to the AIDA on the AWS.

5.2.7. API

AIDA on AWS also offers an API as the interface for the Lakehouse. You can configure it in the same way as in the AIDA on Databricks.

5.3. AIDA on GCP

AIDA implementation can also be done with GCP native data engineering services. All data platform concepts apply also to the GCP services.

Diagram below (Figure 4) shows the architecture of the data platform based on the GCP.

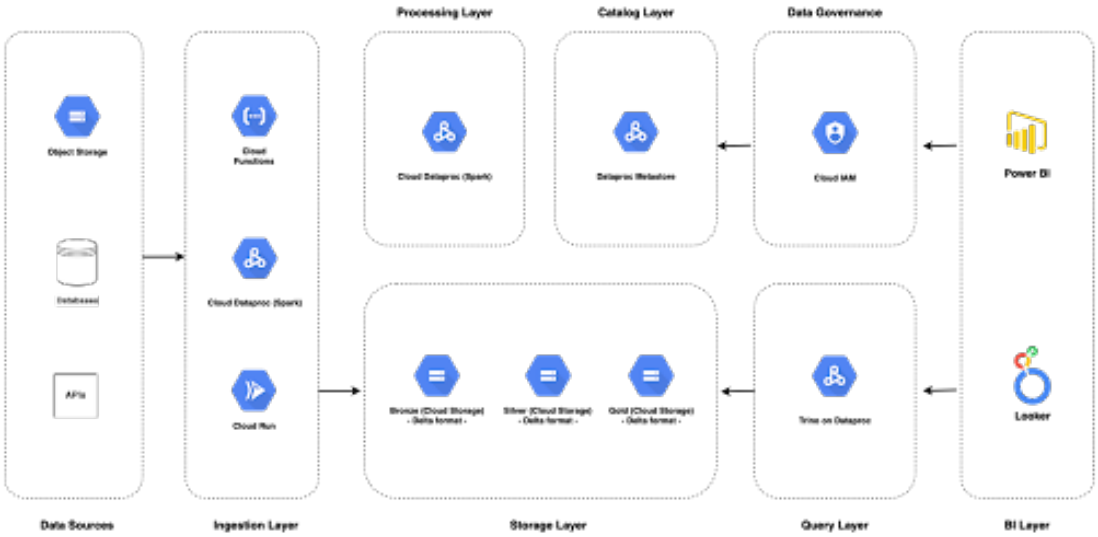


Figure 4. AIDA on GC

5.3.1. The Ingestion Layer

Tooling in this layer depends a lot on the data sources. We can read some data sources using Apache Spark. Apache Spark is managed using GCP Dataproc service.

Data sources that cannot be read using Apache Spark can be ingested using jobs deployed as Cloud Functions or containers on Cloud Run service.

5.3.2. The Processing Layer

In the data processing layer of the data platform on GCP, Apache Spark is the primary tool. Clusters, jobs and orchestration are managed by the GCP Dataproc service.

5.3.3. The Catalog Layer

The Catalog Layer in the GCP AIDA is the **Hive Metastore**. It is an open-source, widely adopted metadata catalog that stores information about the schema and location of the data. Hive Metastore in the GCP ecosystem is managed by GCP Dataproc service.

5.3.4. The Governance Layer

The GCP does not have its own solution for data governance, so the access control is implemented using **GCP IAM** roles and permissions. This approach requires more effort to configure and enable data governance compared to the Databricks approach.

06 AIDA's results in multiple industries

We've tested AIDA in multiple environments (sports betting, vending machines, advertising, Telco...) with outstanding success.

6.1. Retail platform with over 1.000 IoT devices and poor analytics

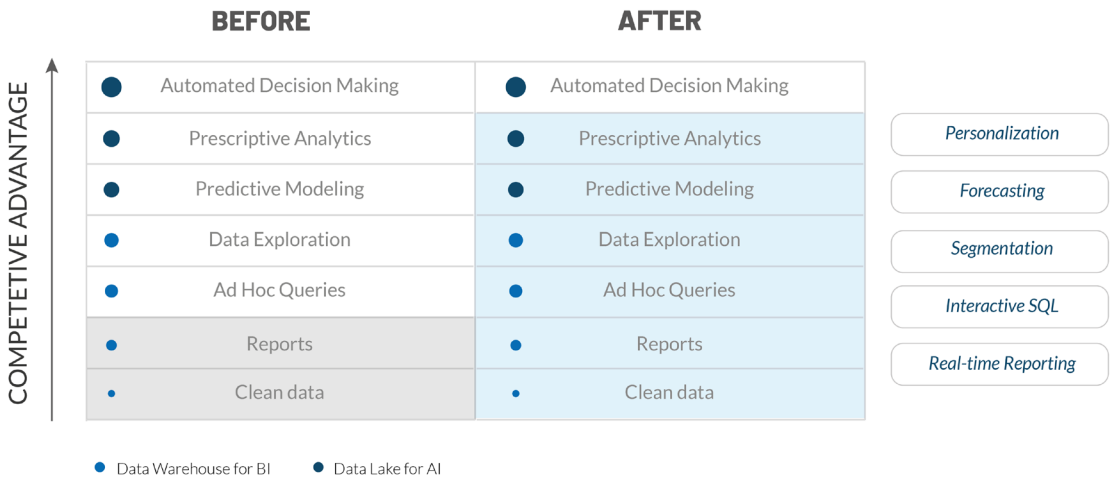


Figure 5. Before and after

The client lacked an OLAP database and therefore they could not do advanced analytics. They had a reporting system that was unscalable, inefficient, limited, and unable to process historical data. Its functions read the data from the storage and could create reports only for the past 2 or 3 months.

We implemented AIDA, which enabled batch and near real-time reporting and advanced analytics. The implementation led to:

- 30x reduced storage and processing costs,
- reduced time-to-insight from weeks to hours,
- enriched data sets from 3rd party providers, and
- reporting on the complete history.

6.2. Online betting system with multiple data sources, incompatible with data analytics tools

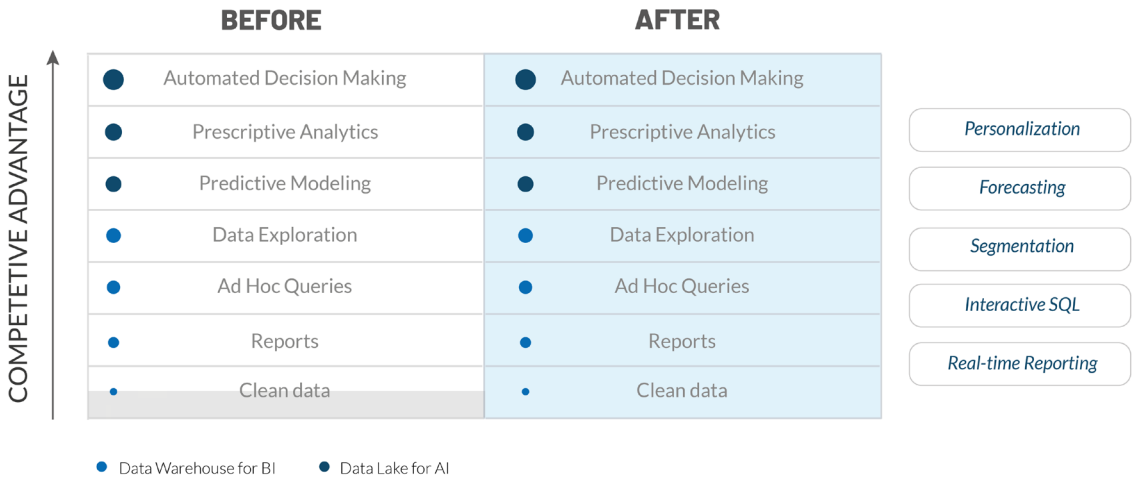


Figure 6. Before and after implementing AIDA

The client had multiple data sources managed by technologies unsuitable for data analytics. They lacked support for stream processing and near real-time insights, and it was an on-premise solution which was difficult to scale.

We implemented AIDA as a cloud-first data platform that provides a unified view of data from heterogeneous systems and enables high-quality, accessible data sets.

This led to **real-time reporting and reducing time-to-insight to minutes**. AIDA provided fast ML/AI experimentation, easy model deployment, and maintenance.

All data teams gained **full access** to data.

07

Interested in AIDA? Here's how to get started.

Even though it promises savings in both time, resources and nerves, implementation of innovative solutions still causes doubts and hesitations.

Mostly because it's unclear what the process is like or the overall impact, how it's measured… There's no room for errors.

Smart Cat takes great pride in having a strong **Business Analysis department** that makes sure all the elements are in the right place (your needs, wishes, expectations and available resources) before we start typing a single line of code.

In a nutshell, here's the process >

1. **Impact Mapping** - through workshops, we analyze your needs, translate requirements into technical points, and draw a map to our solution.

It's a new solution. It takes time to gain trust. We respect that.

2. **AIDA MVP Implementation** - if you approve the map, we set up the infrastructure, integrate **one data source** and prepare data to get valuable insights.

Our goal is to provide you with a quickest time-to-value tool. There's no better way than to give you insights, based on your data, **in the first month already**.

3. **Data source integrations** - continues integration of other data sources and data processing for teams.

Turning an MVP into a full-fledged infrastructure for your organization.

Schedule a demo with our Business Analyst.

Ana Obradović
Business Analyst
ana.obradovic@smartcat.io

Or visit **our website** and learn more about our company, clients, and services.

Thank you for taking the time to read this whitepaper. We look forward to talking with you and helping you streamline data access and insight gathering in your organization.

smartcat.io | Železnička 30, 21000 Novi Sad | +381 (0)69 54 04 007