Eiko

Analytical Platform As a Service

Siavoush Mohammadi | @siavoushm

2020-01-02

ABSTRACT / EXECUTIVE SUMMARY

Analysts, marketers, managers and even scientists of all domains are bathing in data, or rather, as it is in an increasing number of cases, drowning in data. With the eve of internet of things, powerful sensors measuring all sorts of phenomena, super computers running vast simulations of the macro behaviour of the universe, data production is huge, but even smaller organisations are struggling with transforming data to insights.

Organizations are increasingly spending time and resources on building expensive bespoke analytical platforms designed to extract, transform and refine data rather than actually performing any sort of analysis or reporting. Instead, solving exactly the same problems over and over again, by either employing huge numbers of data engineers or forcing the analysts to work extensively with data transformation and refinement.

To solve this we have designed and built an automated, scalable, fast and easy to use Platform As A Service to cater for all of the needed infrastructure, enabling the expert to be experts instead of data engineers. No more need to keep track of updates of specific components, or employing a huge IT department to take care of servers and the bespoke software. This is a Platform As A Service that scales with your business, regardless of the size of it, enabling you to leverage all of the data in your organization from one place with the latest analytical and reporting tools, without being locked-in to a specific vendor or maneuvering the complex big data ecosystem.

TABLE OF CONTENTS

PROBLEM STATEMENT

In a world where everything is producing data, being on top of your data becomes a key factor to success. Historically, data warehouse and big data solutions have been expensive to build and maintain. A large organization usually ends up in struggling to keep costs down and are bogged down by ever slower deliveries, while a small organization can't even afford to develop an analytical platform to begin with and thus loses an important edge compared to bigger competitors.

Those who have chosen to go down the big data path on their own usually quickly realizes the complexity of the Hadoop eco-system and are faced with a myriad of technical decisions that can have a huge impact on everything from performance to complexity, which translates to time and costs. The very same decisions also box the organization into a very specific technology, you become vendor dependent to a high degree.

Another common challenge facing organizations that have made an investment into an analytical platform is scalability and this is not a one direction type of problem as many would think. As your business grows and changes, so must your analytical capabilities without it becoming a huge migration project and development of new platforms.

If you on the other hand don't develop an analytical platform, your organization and employees will find their own solutions to find insights about your business. This can result in several ad-hoc types of solutions, ranging from excels to BI tool installations, and even embryos to data warehouses or big data systems. Not uncommon that these solutions exist in a specific person's laptop, which was a great idea at first, solving a pressing need temporarily (or so you thought) but now poses a great risk for your business, not only because that person might change jobs but more importantly from security and legal privacy requirements like GDPR. Not to mention that you now have no visibility on what data actually is available and what you potentially could do with it.

The solution is never only technology, but also how you work analytically in an agile way. Technology needs to be an enabler, not a blocker.

The pain points for an organisation are:

- High initial development costs
- High complexity of developed systems
- High maintenance costs
- Lock-in effects by technological choices
- Scalability issues
- Agility issues
- Scattered data, i.e. no single point for your data

- Security, privacy and legal requirements (like GDPR)
- Data management

But what if you could get a hybrid big data warehouse, that was scalable, easy to use and have no need for a huge IT department to take care of it? What if the same platform could be ready to use within hours instead of months? What if there was a platform that enabled you to work truly agile with analytics as well as reporting?

SOLUTION

Becoming analytical, is never only about what technology or tools you aquire, but also how you work with them. Technology should be an enabler, therefore Eiko is designed to be a modular and managed end-to-end Big-Data solution so you can focus on what's important, your information and business. Eiko is drawing on over 10 years of Data Warehousing, Big Data, Devops and Fullstack experience of a team of IT professionals. The strength and uniqueness of the solution comes from the in-depth knowledge of our cross-functional team.

Technology: The Cloud Big Data Analytical Platform

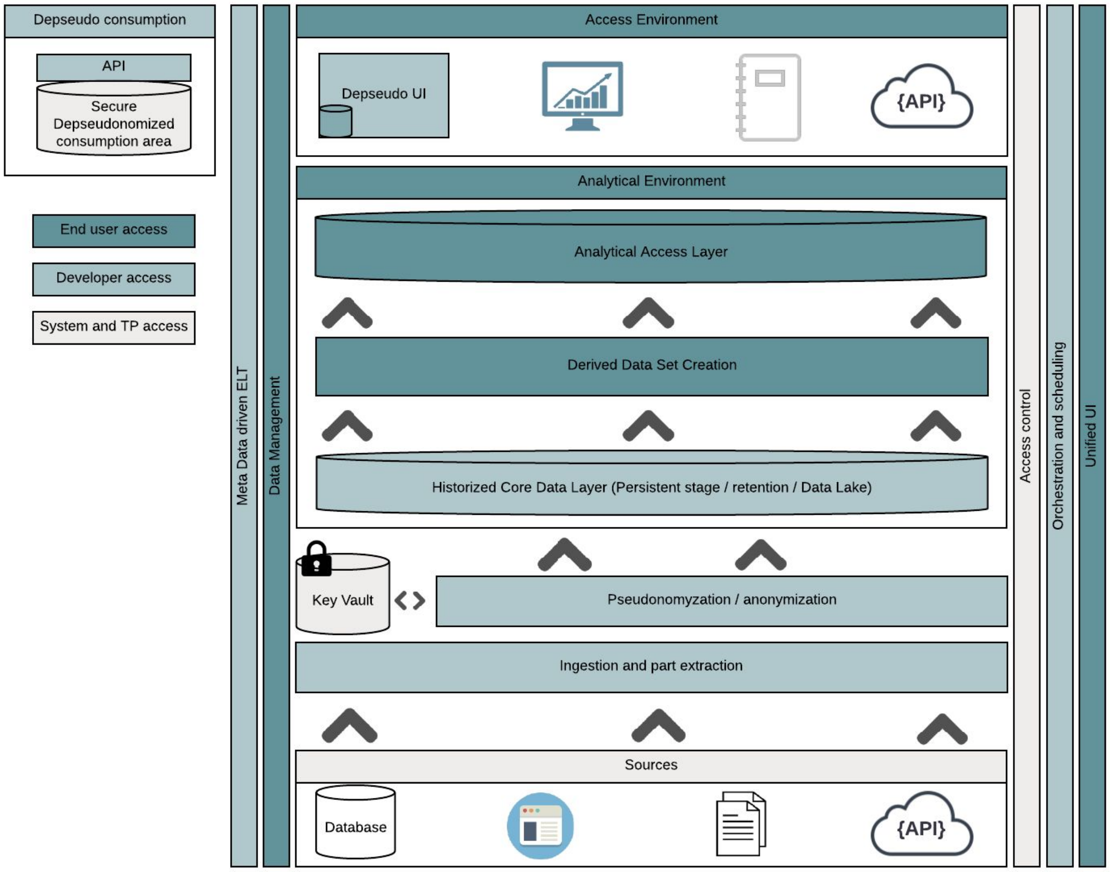


Figure 1 - Architecture of the Eiko platform, the entire platform is a cloud first solution but can be deployed to on-premises solutions also. This architecture enables you an entire analytical platform as a service, working out of the box.

Data Lake and automated DWH working out of the box.

Eiko is a managed Data Lake & Automated Data Warehouse solution, making life easier for the managers and data scientists alike. By being a managed solution, the end-user does not need to worry about dependencies, version management and updates.

Modular by design, Eiko makes it easy to customize your analytical cloud according to your needs. One of the hard things with Big-Data is the size of the ecosystem and the inter-dependencies of the software packages. By managing the configuration of the different combinations of software and testing them we make sure that the system works well when integrated.

To power all of this we have selected Ansible, IT automation tool. This allows us to quickly set up and manage a Big-Data cloud whether the deployment is in a public cloud or on-premise.

Ingestion and privacy by design

One of the challenges many big data and analytical platform implementations are facing are how to ingest data in a standardized and hassle free way, but the solutions are ampel in the open source community, you just need to figure out how to standardize and utilize them in the best way, not to mention make it work with privacy concerns.

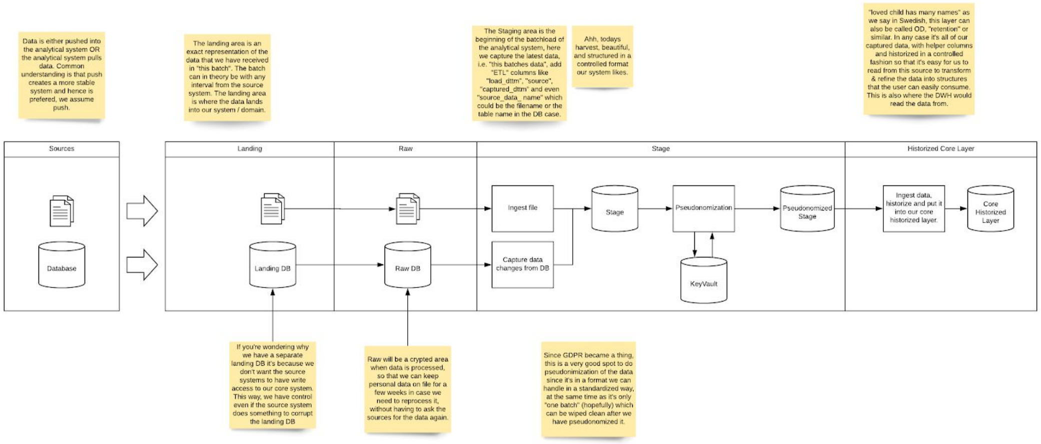
As shown in the architectural overview (figure 1), Eiko supports a wide variety of sources, files (both flat, nested/hierarchical structures and binary), databases, APIs and can even crawl the web and pull information. When data is ingested to the staging area, it undergoes several transformations:

1. Normalization
2. Validation
3. Part extract

then the private data is pseudonymised and stored in Vault by Hashicorp, a component used by many big players in finance to handle sensitive data. The staging and ingestion process is metadata driven with minimal configuration required to start the ingestion.

Ingestion capabilities: The parameter driven ingestion module helps us to ingest data from various sources, examples are listed below :

1) SFTP (File Pull from client server)
2) File Push (Client provides the file to the object store)
3) API
4) Database ingestion



Data is either pushed into the analytical system OR the analytical system pulls data. Common understanding is that push creates a more stable system and hence is prefered, we assume push.

The landing area is an exact representation of the data that we have received in "this batch". The batch can in theory be with any interval from the source system. The landing area is where the data lands into our system / domain.

The Staging area is the beginning of the batchload of the analytical system, here we capture the latest data, i.e. "this batches data", add "ETL" columns like "load_dttm", "source", "captured_dttm" and even "source_data_ name" which could be the filename or the table name in the DB case.

Ahh, todays harvest, beautiful, and structured in a controlled format our system likes.

"loved child has many names" as we say in Swedish, this layer can also be called OD, "retention" or similar. In any case it's all of our captured data, with helper columns and historized in a controlled fashion so that it's easy for us to read from this source to transform & refine the data into structures that the user can easily consume. This is also where the DWH would read the data from.

If you're wondering why we have a separate landing DB it's because we don't want the source systems to have write access to our core system. This way, we have control even if the source system does something to corrupt the landing DB

Raw will be a crypted area when data is processed, so that we can keep personal data on file for a few weeks in case we need to reprocess it, without having to ask the sources for the data again.

Since GDPR became a thing, this is a very good spot to do pseudonimization of the data since it's in a format we can handle in a standardized way, at the same time as it's only "one batch" (hopefully) which can be wiped clean after we have pseudonomized it.

File types Handled:

1) Flat Files
2) Nested Files
3) Binary Files

Gateway to Spark: This section is also driven by the same metadata with bare minimum configuration requirements. Files are always landing in our landing area that we facilitate through minIO (the open source version of Amazon S3).

1) File column names and types are provided and data is binded to schema and put into Spark and HDFS.
2) Each line of data in HDFS will have the source file name, ingestion date and dataset name.
3) Hadoop has limited ability for processing small files, Eiko on the other hand has the ability to merge all data (again metadata driven) according to properties (size and compression).
4) Users also have the ability to keep a backup of data (According to EU regulation, for 14 days) through the archival component (metadata driven) which is generally used to re-process the data in case any process has failed in between. This data store is encrypted and cannot be accessed except if you have the right access rights to do so.
5) Each source has different retention policies, Eiko also has a retention component (metadata driven) to enable your organization to be be GDPR compliant as well as freeing up unnecessary space.

The data layers in HDFS & Spark:

Raw Layer: Which contains the original file.

Core Layer: The data is parsed according to the schema and type conversions applied.

Derived Layer: User might be interested in having some columns out of the entire file. Users can specify the columns required and the derived layer will be automatically created.

One place for all of your data

For any growing organization, increasing data volumes is a big challenge, at the same time just dumping data uncontrolled in a lake quickly turns it into a swamp. Here Eiko can be a saviour with standardized ingestion pipelines integrated with data management practices and technologies. The core data store visible in the above figure uses Hadoop Core API, Spark Core API, Kafka (For streaming purpose), Sqoop, Hive and minIO.

Data management

The purpose of the data management (DM) component in Eiko is to ensure that we always keep track of what data we are loading into the system, why we are loading it, and where we are actually using it. The "where" can mean that the data has been transformed in several steps from ingestion to actual use, but we keep track of it.

All data also needs to be classified as to be part of a subject area, with a clear business definition, and each subject area needs to have a data owner. From a privacy and GDPR perspective this is of utmost importance, since this is where you define if data is personal, the definition of it and purpose of it.

Data Management is more about how you work with data rather than technology. The open source platform we have chosen to facilitate Eikos DM is Apache Atlas, which covers all of the needs.

Visualization, reporting and dashboarding

We are agnostic when it comes to your choice of visualization tool. If your organization prefers Power BI, Tableau, or some other tool, Eiko kan supply the data, both from the Data Lake and the DWH environment. To make sure that the Access Layer has good performance, we are working with Presto DB, a database designed for distributed computing. This way, your visualization tools can make direct queries to the database and still have good performance.

Analytics, Machine Learning and AI

As more and more data is available in the world today, more businesses are adopting advanced analytics with support from big data platform for mission critical decision making. Typically, data scientists first build the predictive models, and then businesses use these models to make decisions . However to get the best out of the models one needs to deploy these models in a production environment and consume them for predictive actions. But the process of operationalizing can be tedious at times and here Eiko provides a very straightforward way of operationalizing the ML pipelines.

- The whole application is designed in modules to break the complex logic of the whole application into more manageable chunks and allow the product to be more agile.
- The variables are organised into different areas. Depending upon the context these variables can be reused.
- The actual runtime environment is separated from the input variables which will allow for code reuse.

- The training only depends on the raw variables: label and all features are extracted from the raw variables.
- The scoring only depends on the dataset used for scoring and is separated from the training module.
- The whole architecture together is complex but as we have broken it down into smaller modules it becomes easy to maintain and to handle them

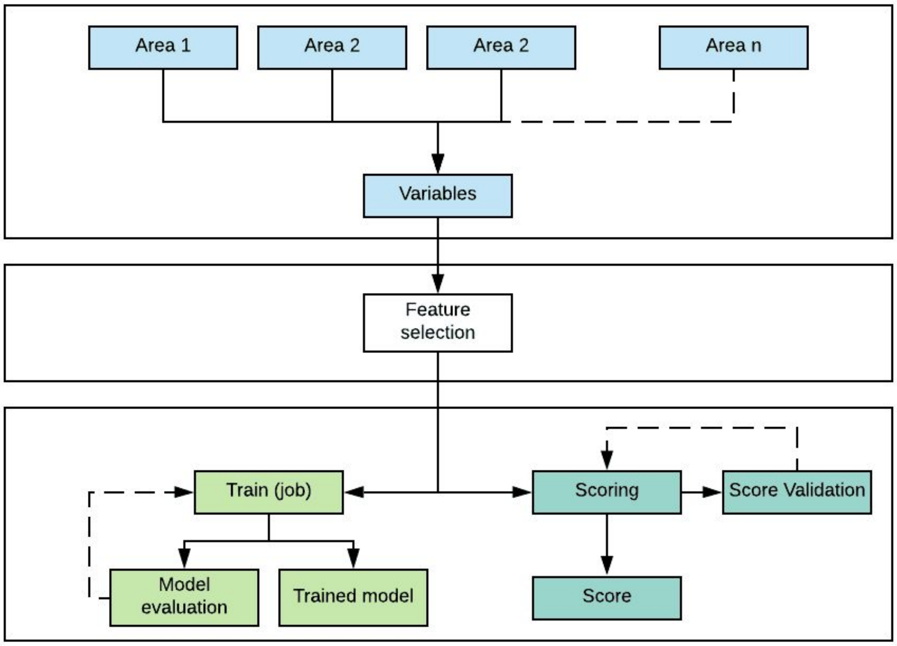


Figure 2: Machine learning operationalization application architecture.

Of course, any BI or reporting tool can be integrated and used on top of the Eiko access layer for easy KPI measuring and reporting.

CONCLUSION

Why Eiko? Because it actually works out of the box, scales both up and down without expensive migrations, is a managed platform as a service, essentially analytics on a tap, where how you work with data in an effective way is deeply rooted in the overall design of the platform and the way you work with the data.

By using Eiko, you get analytics as a service that relieves your organization of

- High initial development costs
- High complexity of developed systems
- High maintenance costs
- Lock-in effects by technological choices

- Scalability issues
- Agility issues
- Single point for your data
- Security, privacy and legal requirements like GDPR
- Data management

All of this has one single purpose, so that you can focus on your business, to create value for yourself and your customers.

FOR MORE INFORMATION

For more information, please contact siavoush@eiko.se