# VATBox: Creating a Data Pipeline with High-Level Security Measures

## Case Study in Data & Analytics







Version 1.0

June 2020


 972-3-9222751 | 972-3-9222907 | www.comm-it.com | 20, Magshimim St. Petah Tikva, Israel POB 7794

## 1. The Customer

VATBox is a world-renowned technological leader in the VAT/GST recovery domain. Their solution fully automates Foreign VAT recovery and ensures governance through one, centralized VAT process across all business entities. In the ever-progressively belt-tightening global economy, financial decision-makers are asking tough questions about their VAT recovery efforts. VATBox provides the visibility and answers needed to understand spend and potential savings. VATBox enforces a single, global process across all entities, using automation and built in domain expertise to improve compliance and significantly increase returns. Website: https://vatbox.com/

## 2. Problem Statement/Definition: Infrastructure for the Unknown

VATBox wished to shift from a data processing model that was based on independent servers and services where each microservice wrote to its own database, some of which were MySQL, some MongoDB, some were other technologies. As a result, they could not create organizational BI. Each DB had part of the data, with discrepancies between them. They did not know as an organization what their data really was. The inability to focus their analysis was stalling them as a business.

Their whole business logic and perspective is one of managing multiple microservices, each with its set of unique rules and processes. Their competitive business advantage was built on the speed in which they could translate their microservices into operations. They had tried batching the new rules that went along with new microservices, but this turned out to be timely and tied up their R&D and IT staff for hours on end, trying to adapt each microservice to fit into the classic ETL.

# 3. Proposed Solution & Architecture

## 3.1 General

CommIT architected their entire dataflow from scratch by utilizing microservices per specific real-time needs and by assigning storage based on the functions required. In other words, instead of focusing on ETL as the centerpiece, we focused on the microservices themselves and leveraged them by producing ETL-processes in AWS Lambda Functions. Each of these ETL processes services only a predetermined set of microservices. In addition, to maintain a single source of truth, all the data is saved as documents into MongoDB on AWS.

## 3.2 Architecture Principal Guidelines

1) There can be several strategy planners.
2) Same stages can be called by different strategy planners.
3) All relevant info except Large Binary Objects (LOBs) is sent from the strategy planner to the stages via SQS message. Any non-LOB sent back from the stages to the strategy planner is also sent via SQS message.
4) The changes in the json files will be marked so when persisting to IDB it will overwrite only the fields that were changed.
5) The LOB info is communicated via REST access to the synchronic IDB Data Access Layer.
6) The strategy planners can override the same Invoice potentially causing data loss, but this situation should be planned. No locking will be implemented in processing of the data.

# 4. Benefits/Outcomes

- **Improved BI:** Utilizing Amazon Redshift for DWH, BI and interactive reporting, the quality and performance improved immensely in comparison to that which they had before (MSSQL).

- **Data Driven:** Utilizing Redshift's columnar structure and analytic focus, they could make better data-driven decisions.
- **Aggregated Queries:** Redshift was the exact solution to provide this capability, which was the very reason the client needed this solution, replacing the slow or unfeasible solutions they had tried to utilize before.
- **Cost:** The costs of running the old system started at $100,000 annually, whereas the new system operates runs on around $1,000.
- **Data Collection:** The new architecture replaced the old, expensive Informatica platform, which collected data in 50 separate services. The speed, efficiency and reliability of the new system far outweighed the old system.
- **Centralized Source of Truth:** It was a complex process to add microservices. The new architecture provides one central single source of truth DB (MongoDB), especially in comparison to the old way of working via dozens of MSSQL servers with pieces of data each.
- **Simplicity:** The ETL required to retrieve information from MongoDB and to shape it and load it to Redshift is a simple, flexible and low-cost mechanism.
- **Ease of Expansion**: Most importantly, since deploying the new AWS-based data and analytics solution, VATBox has added more microservices easily and swiftly. They no longer have to update systems and data manually, since the new automated procedures do this for them, in a secure and consistent manner. This has allowed them to offer reliable new services to the market, increasing their profits and market share considerably.

## 5. Lessons Learned

An interesting point we learned in the process was the difference of data usage, and how to match that to the correct data and analytics services and tools. At the end of the process, managers are given the ability to present the summaries of the data collected, but also can

modify the information. They can approve the sums, approve returns, exchange rates and more. This means they needed updatable reports. When examining which database platform would be best we saw that Aurora far outweighs Redshift.

That is why, for the purpose of updatable reports, we added a component. The reports were replicated from Redshift to Aurora Postgres, and then the serving layer was constructed inside the latter. Changes could be made to the report, and then updated in the central source of truth – the MongoDB document store called IDB. From there, the Redshift data is corrected as well, which means that the data at the end of the process is the same across the board.

This out-of-the-box thinking has taught us that each need can have a service that matches it, which will be cost-efficient and business-efficient as well.