WHITE PAPER

French Hospital Uses Trusted Analytics Platform to Predict Emergency Department Visits and Hospital Admissions

Trusted Analytics Platform (TAP) is a collaborative environment for creating advanced analytics and applications to help hospitals improve patient care and resource allocation.

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

www.aphp.fr

Kyle Ambert
Data Scientist and Health Analytics Technical Lead, Intel Corp., PhD

Sébastien Beaune
Emergency Department Director, AP-HP Ambroise Paré., MD, PhD

Adel Chaibi
Application Developer, Intel Corp.

Luc Briard
Data Scientist, Intel Corp., MBA

Amit Bhattacharjee
Data Scientist, Intel Corp., MSc

Venkatesh Bharadwaj
Data Scientist, Intel Corp., MS

Krishna Sumanth
Data Scientist, Intel Corp., MS

Kathleen Crowe
Director of Data Science, Intel Corp., PhD

Overview

For hospital administrators, predicting the number of patient visits to emergency departments, along with their admission rates, is critical for optimizing resources at all levels of staff. Ultimately, this reduces wait times in emergency departments and improves the quality of patient care.

Intel and the Assistance Publique-Hôpitaux de Paris (AP-HP), the largest university hospital in Europe, worked together to build a cloud-based solution for predicting the expected number of patient visits and hospital admissions using advanced data science methodologies and the Trusted Analytics Platform (TAP). TAP is an open source platform that accelerates the creation of applications driven by big data analytics. Using data from four emergency departments within AP-HP, data scientists from Intel and medical experts from AP-HP evaluated three different approaches to time series analytics, optimizing model parameters and identifying the best predictive features to include in each. The team selected an Autoregressive Integrated Moving Average with Exogenous Input (ARIMAX) approach that proved to be simultaneously accurate, scalable, and easily adaptable to the needs of both data scientists and hospital staff.

The team moved into the model optimization phase of the project, using such metrics as the Akaike Information Criterion, or AIC, to explore which features to include in the model to balance accuracy and complexity. The team also developed an Apache* Spark-based implementation of the ARIMAX algorithm to take advantage of the speed and scalability of TAP's distributed processing infrastructure.

The data science team then implemented the data model into TAP for efficient cloud-based computation and to streamline workflows with application developers who created a secure, browser-based user interface to interact with the results. This prototype interface, which is undergoing testing at AP-HP hospitals, will enable hospital administrators to view 15-day predictions of emergency department visits and hospital admissions to optimize staffing levels based on anticipated needs. In the future, AP-HP plans to leverage TAP's highly-scalable environment to process large data sets to increase the accuracy predictions, and to investigate other healthcare challenges that could be addressed through the analysis of big data.

Business Drivers

Overcrowding is a growing problem in hospitals around the world. It results in backups in the emergency room, with patients occasionally being turned away from emergency departments due to lack of space, beds, or staff to treat them.

Predicting the utilization of their emergency departments is a major priority for administrators at AP-HP, both as a means for reducing patient wait times and for improving patient care. Inability to accurately plan for future patient loads is a cause of frustration and concern, as some emergency rooms in AP-HP may be overcrowded while others may be underutilized. In addition, patients may not have sufficient information about specialty clinics within the network to ensure they go to the most appropriate emergency department for specific ailments or injuries. These challenges make it difficult for administrators to effectively allocate staff and other resources, because they are unable to accurately anticipate the need for beds both in emergency departments and in other parts of the hospital.

Emergency Departments with Rich Datasets

In 2003, an intense heat wave hit France, resulting in the death of dozens of citizens, in part due to inadequate emergency health facilities to deal with the crisis. The French government responded by making major investments in new emergency departments, including new electronic information systems. With a decade's worth of both clinical and operational data within its emergency department networks, AP-HP has a rich dataset with the potential to use as a basis for meaningful big data analytics solutions for addressing a range of challenges in healthcare. In addition, clinicians and data scientists at AP-HP were aware of recent studies suggesting correlations among such external factors as weather, holidays, and flu incidence and increased emergency room visits.

AP-HP began discussions of how to use its data to support development of an analytics and machine learning solution that could help predict emergency department visits and hospital admissions. They wished to incorporate additional data sources into a predictive model with their own data, to evaluate whether including external factors would contribute to a greater accuracy in visit- and admission-rate projections. AP-HP's initial goal was to develop a web interface that would allow hospital administrators to view predictions of future emergency department visits and hospital admissions. A future goal was to offer a public-facing version of the interface that would allow incoming patients to view anticipated wait times at all of the network emergency departments and to evaluate the medical specialties at various hospitals before determining which AP-HP facility to visit.

"The primary consideration in the development of our data science project with Intel was to improve patient care and outcomes," said Raphael Beaufret, the head of web innovation in the AP-HP CIO's data department. "If patients had access to accurate real-time data on wait times at all AP-HP emergency departments, and information about the availability of specialized care opportunities at various locations, some of them could make more informed decisions about where they seek care."

Project Overview

"We worked closely with AP-HP to build a prototype cloud-based solution designed to predict the expected number of patient visits and admissions over a moving 15-day window. This proof-of-concept (POC) used retrospective data to build a predictive model of patient visits for four of the French hospitals in AP-HP," said Kyle Ambert, senior data scientist and technical lead for Health & Life Science, Intel Corp.

Also contributing to the collaboration was ETALAB, a bureau of the French government responsible for the direction of France's open data initiatives. The AP-HP project was developed at Teratec, a Paris-area campus, founded by a consortium of government agencies, academic institutions, and private companies, including Intel.

Data Wrangling and Preparation

France has strict data privacy laws, which prevented AP-HP from providing Intel with its source data directly.

Because ETALAB had permission to work with the protected health information contained in the raw data, it was responsible for managing the full anonymization of the dataset, removing any information that could lead to the identification of a patient. The dataset provided to Intel, which contained data on more than 470,000 patient encounters, included the hospital site of the encounter, whether the patient was admitted to the emergency department or later to the hospital, and the date and times of patient movements in or out of the emergency department.

Given the limitations brought on by dealing with anonymized records, data scientists from AP-HP and Intel were restricted in the queries that could be performed on the data. However, by including environmental data, such as weather and flu rates from publically available data sources, they could explore the usefulness of adding these external data sources to provide more context to the patient information.

Phase One: Modeling

After the full dataset was prepared and assembled by ETALAB, the project moved forward in two phases. The first involved the bulk of the data science research, along with the initial model evaluations. During this phase, the Intel and AP-HP teams cleaned and sampled the data as well as surveyed a variety of possible methods for predicting hour-by-hour visits and admissions at each site.

Though the team's primary goal at this point was determining the time series algorithm that most accurately tracked the peaks and valleys of the observed data from AP-HP, there were also important secondary goals. The selected analytical model would need to scale for the analysis of much larger data sets than those used in the POC, and easily adapt to a distributed computing framework. The model would also need to be readily interpretable to a non-statistical group of users, namely clinicians.

The team evaluated three time-series analysis approaches for deployment with the AP-HP data. For each method, the team optimized model parameters, as well as identified the best predictive features to include in each.

Phase Two: Selecting and Training the Model

After evaluating the three time series analysis models, the data science team from Intel chose the ARIMAX approach, based on its ability to address the requirements for accuracy, scalability, and versatility. The team then tuned the model for production on TAP.

ARIMAX is an algorithm used frequently in the financial sector, where it's employed for predicting changes in stock, commodity, and other financial market value. It provides a standard approach for expressing a linear regression model in which predictive variables are used to predict some future outcome. ARIMAX takes into account previously-observed values and incorporates them into the predictive model.

With an optimized model and a final dataset (which included the anonymized data from AP-HP, as well as a number of external, open datasets, including weekly reported flu cases throughout France, maximum and minimum daily temperatures recorded at Charles De Gaulle airport, and holiday and post-holiday indicator variables), the team was ready to deploy the model on TAP where big data analysis could quickly be performed and results easily shared with application developers.

Solution Details

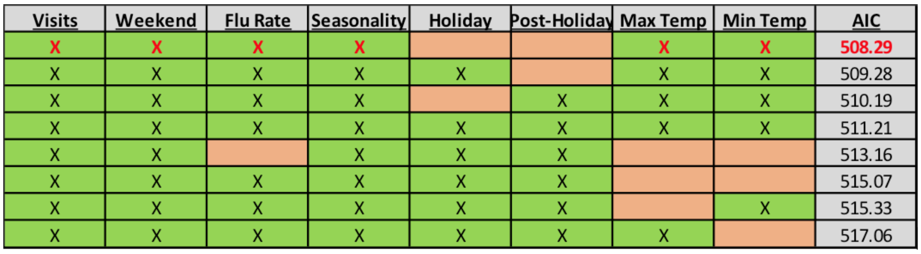
The data science team evaluated the final dataset and the ARIMAX model with respect to the AIC, a metric for quantifying the tradeoff between model accuracy and model complexity. Generally speaking, data analysts aim to find a model that leads to the greatest accuracy for a given set of data. However, to focus simply on increasing the model's accuracy by adding an abundance of predictive variables or over-training the model for a data subset can lead to overly complex models that aren't able to generalize to new data. AIC quantifies the accuracy of a model for a given set of training data, along with the likelihood that the model will generalize to unseen data.

The team performed a traditional model comparison of the different parameterizations for ARIMAX to identify and fix the correct parameter settings to minimize the objective function (the AIC), then evaluated the external feature combinations to similarly minimize AIC. By running successive tests and adding and removing variables, the team arrived at a model that yielded the greatest accuracy with the least complexity.

The results of the model exploration phase, using the AIC metric, indicated that the best-performing models (see Figure 1 – a lower AIC number is better) did not utilize all the explanatory variables that could be logically expected to be useful as predictors. The optimal combination of features for this dataset were the estimated number of visits, a weekend indicator variable, flu rates, seasonality, and the daily maximum and minimum temperatures at Charles De Gaulle airport. Notably, models using holiday information did not perform as well, indicating that these data were not predictive for emergency department visits and hospitalization for this population.

To illustrate the contribution that a single feature such as seasonality could make in the model, the team compared the predicted and actual admission rates for the model with and without seasonality. Additionally, this comparison enabled visualization of the predicted and observed admission rates (see Figure 2).The models generally followed the patterns of the observed data well, with an overall variance of around 5 percent. However, the accuracy of the model speaks to the inherent difficulty of predicting actualities that are triggered by accidents and unforeseen events.

The data science team deployed a distributed implementation of the ARIMAX time series algorithm on TAP, to take advantage of TAP's capacity for massive scalability and fast processing. The ARIMAX algorithm utilizes the Apache Spark Resilient Distributed Datasets (RDD) paradigm, which keeps RDDs in persistent memory for iterative processing to minimize the amount of time spent moving data into memory for repeated computations. The data is distributed to different compute nodes in the cluster for processing, then reduced back into memory on the control node, with no need to reengineer the code for scalability as data size increases.



Visits	Weekend	Flu Rate	Seasonality	Holiday	Post-Holiday	Max Temp	Min Temp	AIC
X	X	X	X			X	X	508.29
X	X	X	X	X		X	X	509.28
X	X	X	X		X	X	X	510.19
X	X	X	X	X	X	X	X	511.21
X	X		X	X	X			513.16
X	X	X	X	X	X			515.07
X	X	X	X	X	X		X	515.33
X	X	X	X	X	X	X		517.06

Figure 1. Impact of the combination of datasets within AIC (lower AIC number is better).

TAP will ensure that the optimized model will scale to support datasets far larger than the current data (derived from four hospitals), as the ultimate goal is to analyze data from all 44 locations. Not only is TAP designed for fast processing of large datasets and for massive scalability, but it also includes ability to incorporate new algorithms, tools, and custom features as requirements evolve.

Contributions to the Open Source Community

Another outcome of the team's work on the AP-HP project was its contribution of the fully-developed ARIMAX time series algorithm to spark-ts, a library for time series analysis on Apache Spark, to the open source community.

Browser-Based Application

In TAP, the data science team built a browser-based user interface for hospital administrators, physicians and medical staff to easily visualize forecasted near-term hospital patient loads and plan resource allocation.

In its current implementation, the application enables hospital employees to view the past 15 days of reported data for emergency department visits and hospital admissions, and to view predicted visits and admissions for the subsequent 15 days.

Because the application was also developed in TAP, data scientists and application developers, from Intel and AP-HP were able to work together closely during the development. Figures 3 and 4 illustrate the application user interface.

TAP Benfits with the AP-HP Proof-of Concept

For the Intel and AP-HP analytics team, scalability was one of the primary benefits of using TAP. Once the code and the application were optimized, the solution could scale to process much higher volumes of data. This was particularly important when running successive time series models to

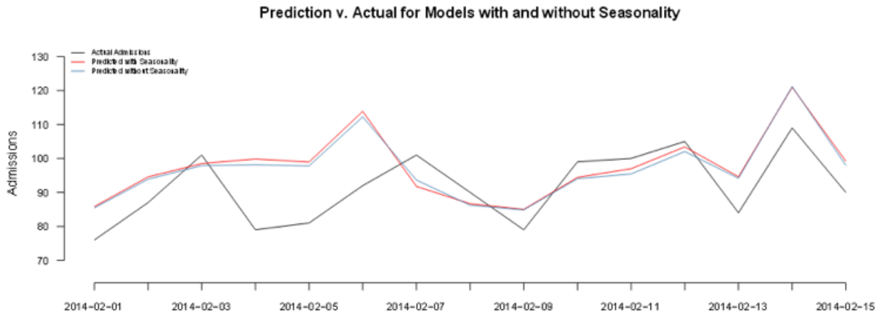


Figure 2. Admission predictions for a model with (red) and without (blue) the seasonality feature, compared with the actual observed rates (black).

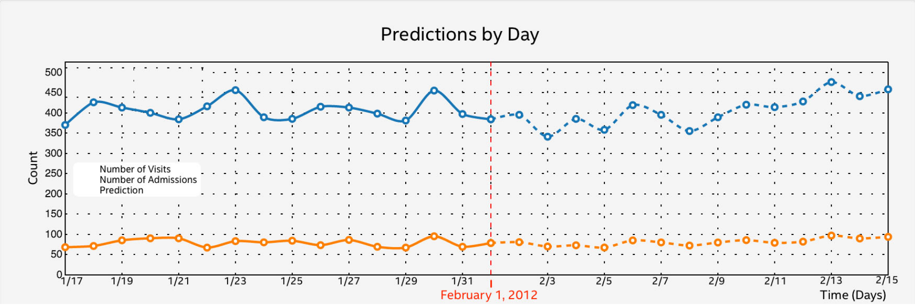


Figure 3. The blue line represents the number of visitors to one of the AP-HP emergency departments, while the orange line represents admissions to the hospital from the emergency department at that location. To the left of the vertical red dotted line, which represents the present day, the data points reflect reported data while the dotted lines to the right are predicted levels of utilization.

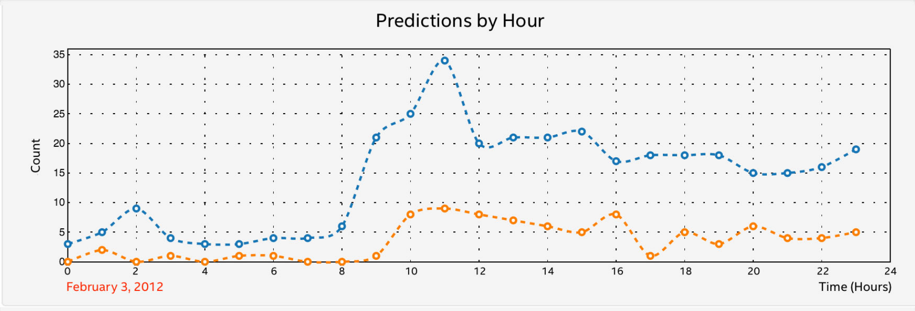


Figure 4. The blue line represents the number of visitors to one of the AP-HP emergency departments, while the orange line represents admissions to the hospital from the emergency department at that location. To the left of the vertical red dotted line, which represents the present day, the data points reflect reported data while the dotted lines to the right are predicted levels of utilization.

provide data for analytical predictions. Forecasting the number of admissions to the hospital for any given day relies on input from the prediction of that same day's number of visitors to the emergency room. TAP's ability to scale for efficiently running these large and complex time series models is important for the effectiveness of the AP-HP solution. TAP's scalability will also be key to the solution's ability to efficiently process datasets from AP-HP's larger data lake, fed by its 44 member hospitals.

Although AP-HP data provided to the Intel team had been anonymized, the TAP platform is optimized for security to ensure that data and processing are protected from end to end. Had AP-HP run its data on an internal cloud-based TAP environment and used its own data scientists to analyze the data, TAP's security features would have satisfied French privacy mandates and de-identification of the data would not have been necessary.

Statistical experimentation on models and algorithms formed a large part of the initial research into the AP-HP solution. TAP's collaborative environment for advanced analytics facilitated the iterative back-and forth between teams of data scientists that enabled faster model development and ensured they performed as intended. In addition, the capabilities of TAP extend into solution development, enabling data scientists and application developers to work together closely in a shared environment to iterate both data and application formats for delivery of an optimal solution.

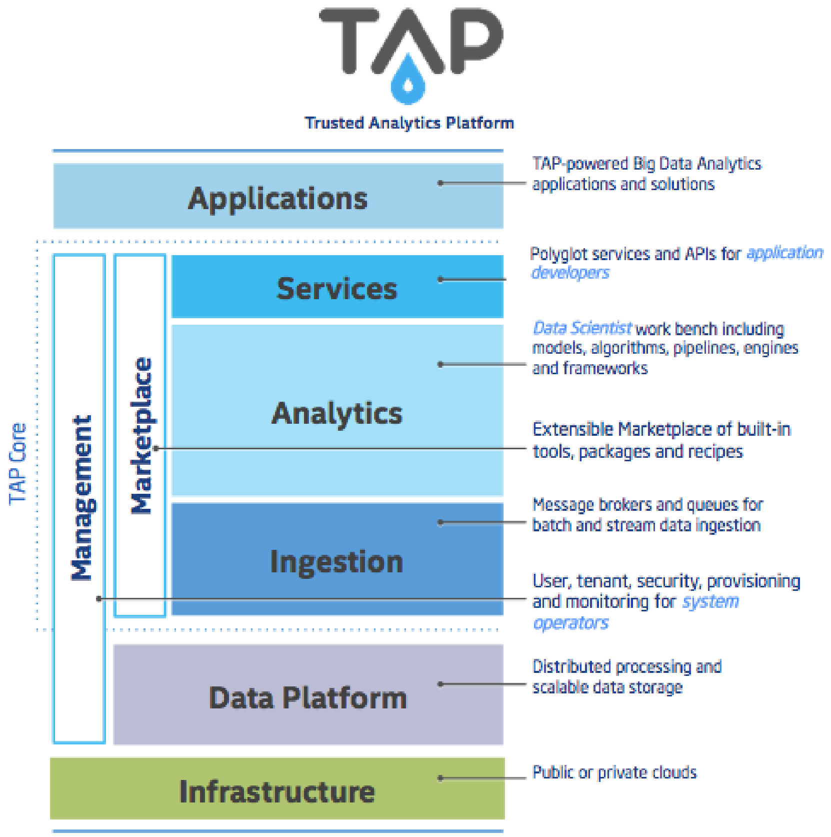
Summary

"We undertook a joint project, with AP-HP, to research and implement a big data analytics-based solution for predicting emergency room visits and hospital admissions in a university hospital in France. Though the project is still in its early phases, the Intel and AP-HP data science teams have succeeded in creating a POC that met AP-HP's goals for the project and points to further opportunities

Trusted Analytics Platform

TAP is open source software optimized for performance and security that accelerates the creation of applications driven by big data analytics. TAP makes it easier for application developers and data scientists — at enterprises, cloud service providers, and system integrators — to collaborate by providing a shared, flexible environment for advanced analytics in public and private clouds. Data scientists get extensible tools, scalable machine learning algorithms, and powerful engines to train and deploy predictive models. Application developers get consistent APIs, services and runtimes to help integrate these models into applications quickly. System operators get an integrated stack that they can easily provision in a cloud infrastructure.

TAP helps lower development costs and reduces time to deploy analytics applications for organizations that want to create custom solutions using big data analytics. Tested by data scientists in various industries, TAP uniquely provides a complete pipeline for graph analysis as well as scalable algorithms and in-memory engines for machine learning. TAP delivers open source software as an integrated platform with hardware enhanced performance and security in every layer.



TAP-powered Big Data Analytics applications and solutions

Polyglot services and APIs for *application developers*

Data Scientist work bench including models, algorithms, pipelines, engines and frameworks

Extensible Marketplace of built-in tools, packages and recipes

Message brokers and queues for batch and stream data ingestion

User, tenant, security, provisioning and monitoring for *system operators*

Distributed processing and scalable data storage

Public or private clouds

Figure 5. The anatomy of TAP.

to collaborate in advancing the role of data analytics in healthcare," said Ambert.

"In addition to helping specific hospitals organize and allocate resources to avoid overcrowding and other emergency room challenges, I see a number of other uses for this solution," said Dr. Saïk URIENS, Unité de Recherche Clinique et CIC Paris Descartes Necker Cochin, APHP, Research Director at Inserm. "Moving forward, our hospital administrators and medical staff should be able to use this predictive analysis, on retrospective big data, not only to estimate the number of emergency department visits (including admissions rate), but also to categorize these visits. For instance, it would be very helpful to have predictive data that showed the proportion of children, adults and elderly persons or men and women, as well as type of medical causes (infection, cardiology, surgery, traumatology, etc.)."

"Seeing the prediction application take advantage of all the data and provide useful and actionable insights has allowed our medical staff to imagine the tremendous benefit it will provide to both the staff and our patients," said Dr. Sébastien Beaune, emergency department director at AP-HP. "Having a better understanding of patient flows at our emergency departments—or even predicting these flows—is absolutely key if we want to improve our quality of care."

The Intel and AP-HP collaboration successfully created a merged dataset that incorporated de-identified data from four AP-HP hospitals with exogenous public datasets that described external conditions in France that might relate to hospital visit and admission rates.

The data science team evaluated a number of time series analysis algorithms to determine which approach optimized requirements for accuracy, scalability, and interpretability by non-data science users. The team chose the time series analysis method ARIMAX and deployed it on TAP, a collaborative, flexible environment for developing advanced analytics solutions. The TAP analytics and development environment was also instrumental in creating a prototype browser interface that displays predicted patient visits and admission rates. The result was an application that makes it simple for hospital administrators and clinicians to better anticipate staffing levels and improve allocation of resources, all with the end goal of improving care delivery for patients within the AP-HP hospital network.

To learn more, visit:

Intel Big Data & Analytics:

https://software.intel.com/bigdata

Trusted Analytics Platform:

http://trustedanalytics.org

AP-HP:

http://www.aphp.fr

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configution No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at www.intel.com. Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725 or by visiting Intel's website at http://www.intel.com/design/literature.htm.

Printed in USA 0516/MD/MCB/PDF Please Recycle 334409-001US