Building Digital Transformation to improve NGCTR design and simulation

Michele Sesana, Alessandro Pietro Bardelli

TXT e-solutions S.p.A. Via Frigia 27, 20126 Milano, Italy
michele.sesana@txtgroup.com, alessandro.bardelli@txtgroup.com

Abstract. New technologies and increasing accessibility of powerful hardware are opening a new world of opportunity to digitalize current processes making them more efficient. The paper will discuss the process of digital transformation of the NGCTR flight test campaign in which the new big data analysis technologies have a central role in the gathering, storing analyses and provision of evidence about what happened during flight. The new approach is enabled thanks to a powerful big data analysis cluster which have been tailored to best support the specific process.

1. Problem Statement

If an airline can predict that when a part is going to fail and to prevent it from happening, extra costs and passenger annoyance during flight can be avoided. Combining flight analytics and sensor data from engines with customer data, airlines can better manage flight disruptions, not to mention missed connections. This is achieved by using data that the aircraft generated during flights through their deep analysis in order to detect possible malfunctions, performing predictive maintenance, anticipating problems before it is raised.

Modern aircrafts are based on a rich set of on-board computers supporting many different tasks, capable of recording hundreds of parameters during flight. This allows not only the investigation of problems occurred during flights, such as accidents or a serious incident, but also provides the opportunity to use the recorded data to predict future aircraft behavior. This can be done with precise analyses of the recorded data, in order to identify possible hazardous behaviors and developing procedures to mitigate the problems before they occur. Because of the enormous amount of data collected during each flight, two are the problems to be faced: define the appropriate infrastructure in terms of HW platform and the most promising software to identify the segments of data that contain useful information. The objective is to extract for a huge amount of data useful information that can support the aircraft manufacturer to better support their work. Traditional data-mining methods are effective on uniform data sets such as flight tracking data or weather. Integrating heterogeneous data sets introduces complexity in data standardization, normalization, and scalability. The variability of underlying data warehouse can be leveraged using big data infrastructure for scalability to identify trends and create actionable information. The massive availability of data requires complex and performing architecture to support deep analysis on data. Furthermore, data can be so huge that an intelligent support shall be provided by the platform itself, that is, the end user shall be reinforced and guided in the analysis by means of an intelligent support. This can be only achieved by means of the...
adoption of novel approach to analyses large amount of data and extract useful information (data mining, machine learning, AI).

The ADMITTED Project has the goal to enable an innovative approach in the development of the NGCTR-TD by implementing methods capable to supporting the optimization of flight campaigns and improving design choices, thanks to new insights coming from the massive data analysis and finding innovative algorithms and correlations. Dimensions of a typical flight test data are 3 Flying aircraft prototypes, an average of 10,000 and a maximum of 30,000 parameters, for each flight condition including 600,000 of total flight conditions and the prototypes have as today about 4,000 flights recorded.

Main Objectives of the project are to:
- define the most appropriate infrastructure to support large amount of data collected during flights test both in term of HW big data platform and SW toolset to support storing, retrieval, data analysis and ii) to define the best approach to extract useful information from the data recorded according to the identified main purposes;
- select the most promising techniques to support data analytics supporting computation of prediction or suggestion based on data;
- the implementation of novel predictive algorithms based on machine learning techniques.

2. Progresses beyond the State of the Art

The Flight Data Recorder (FDR) is used in commercial aircraft to collect and record data from various aircraft sensors. Often it is combined with the Cockpit Voice Recorder (CVR) that is used to record the voices of the pilots and/or camera. However, FDR are used more for accident analysis after a problem is raised. Nowadays, on board computers on modern aircrafts are able to collect only a limited amount of flight information that constraints their use. Current limitations, needs and gaps includes the fact that data are collected on board and processed on ground after flight. This is not a real limitation if the purpose of such data is to perform back office analysis instead of detecting real time indications. Data are recorded on specific mass storage devices and filled in the platform after flight. The amount of data is fairly huge (up to 30000 parameters per flight, 3 flying aircraft prototypes…); Specific platforms are required, both HW and SW to store all data and make them available to algorithms in short time. To retain data in files is not a viable option anymore: too much time is required to properly extract data for elaboration. ADMITTED contribution beyond state of the art is that flight data are collected and injected into the computation cluster after aircraft flight. Data ingestion is done through specific ETL (Extract Transform Load) algorithms automatically reading information from downloaded files. The online communication with the aircraft for online data transfer is not part of this proposal due to the missing availability of adequate infrastructure. However, data ingestion algorithms will be implemented in a way that they can support a future online implementation.

On Data analysis algorithms Big data platforms support the analysts in developing algorithms with a large set of tools. Specific frameworks, such as Spark, have been designed to provide a high-level approach to data analysis. However, this requires a change in the mind set of technicians approaching the new generation of tools. A deep and long learning path is required to be able to get benefit from these new platforms. Currently the traditional tools like Matlab – and even Excel – are no more usable. They are not able to manage such a large amount of data. Matlab is a valuable tool for data analysis, but it can operate on a limited set of data (with respect of the currently available quantity) and may not highlight all the complex signals correlations that may emerge only with a massive analysis. Nevertheless, there is a large number of algorithms developed in the recent past for Matlab
environment. The most important and promising algorithms shall be converted to the new SW platform. Processing high volumes of data takes unrealistically long time, making the real-time processing virtually impossible. Combining heterogeneous data sources requires custom solutions. Real-time overview/report/visualisation generation for high data volumes is mostly impossible. Context of the data is frequently lost during database export and/or data analysis. Queries incorporating expert knowledge and/or statistical/physical formulas are complicated to create and maintain. ADMITTED Contribution beyond state of the art is a set of "traditional" data analysis algorithms that will be provided as basic catalogue (data filtering, FFT, correlation, …). Beyond them a set of algorithms are defined, specifically for flight data analysis, operating on statistics and machine learning. The adoption of a powerful HW platform will make practical the analysis of huge amount of data. Data correlation and comparison among flights from different aircrafts are possible only if supported properly by the underling platform and algorithms that know how to exploit it.

About Aircraft design phase & Flight campaign optimization flight testing is a significant cost contributor to the aircraft production life cycle and is still extensively deployed. Flight test programmes take several years, and more prototypes are built to reduce lead times. Experimental design is foremost applied at an academic level, at early TRL levels, but it is not achieving its full potential in the industrialized flight-testing environment. Strong adherence to safety and certification requirements and generally unchanged circular advisories inhibits the potential improvement of flight test designs. A collection of testing techniques in the area of statistical estimation can increase the quality and productivity of an experiment, leading to a required test point reduction or increased predictive capabilities. With current implementations only a partial comparison is possible, only few sets of parameters can be combined during analysis. This may lead to erroneous considerations or take a considerable amount of time before to obtain the desired values. The ADMITTED implementation will be able to manage large amount of data, as stated in previous topics, but also will provide a considerable help in the understanding of data thanks to the adoption of machine learning techniques. When properly trained, it can support the data scientist in better understanding the characteristics of the new aircraft (NGCTR).The result of flight data of previous developed aircraft can be used to support the design of new aircraft (such as the Next-Generation Civil Tiltrotor). In addition, the use of engineering data of the aircraft under design can be compared with flying data.

3. Findings in the first year of the project

The first year of the project intends to support data scientists during their analyses on flight data according to the following key steps:

- Definition of a suitable computing node, having enough computation power and disk space to support several aircraft data
- Definition of a ready to use set of basic analytic features, to be adopted for data analysis
- Adoption of an intelligent support in data analytics by means of machine learning based approach to specifically address the key issues of the aeronautic domain
- Adoption of novel AI techniques to better support the understanding of flights occurrences. In fact, one of the key features of the proposed platform is the ability to combine heterogeneous data coming from different sources: flight data, time series form sensors, environment conditions, aircraft settings, pilot interaction and other analyses

The first important result achieved is the selection, installation and connection of the enabling infrastructure of the project which will be used by the project during the five years. The enabling infrastructure based on a complex, yet flexible HW/SW combination is composed from HW perspective of:
- A baremetal infrastructure of 4 nodes optimized for storage and cpu-intensive computations
- A Hyperconvergent Nutanix architecture built upon the beforementioned 4 nodes that is configured to expose the user a 6 nodes cluster capable of holding 80Tb of data.
- A single node powerful workstation equipped with two best-in-class V100 Nvidia GPUs for deep learning applications

From a SW/application perspective:
- A 6 node Cloudera cluster configured with all the Hadoop related services (e.g., Spark, Hive, Kudu, Impala, Hue) needed to run ADMITTED ETL and data analysis jobs
- A standalone workstation equipped and configured with a complete python distribution and all the main Deep Learning and Data Science toolboxes and libraries such as Numpy, Scipy, Pandas, Tensorflow, Pyspark.

![ADMITTED Cluster](image)

**Figure 1:** ADMITTED Cluster

On the four computation nodes has been installed and configured the Nutanix distribution in order to guarantee the flexibility of creation and hosting of virtual machines thanks to the Hyper convergent Nutanix Architecture. At the same time the specialised node equipped with GPUs has been configured in a standalone mode so to be able to fully exploit its computational power and devote it entirely to Machine Learning and Deep Learning applications.

Out of the 4 Nutanix nodes, 6 new virtual machines have been then created and configured with the Cloudera Hadoop Distribution. Lot of useful services have been installed to complete the software stack equipment and make the cluster suitable for the different applicative scenarios foreseen in the ADMITTED project. The following figure shows the list of the services deployed on the cluster and the roles assigned to each computational node of the Hadoop Cluster.
At the same time, the Deep Learning workstation has been equipped with a CentOS distribution, has been properly configured in order to access the cluster and exploit it to perform the ETL (Extraction Transformation and Load) operation needed to prepare data for Deep Learning applications. Also, it has been equipped with a rich software and application stack to ease the work of the Data Science team.

The cluster and the Deep Learning Workstation can be accessed via SSH from configured laptops. The following figure provides an overview of the final architecture and shows how the Deep Learning Workstation is also the user entry point for both the Deep Learning and GPU resources and the Big Data and cluster’s ones as well.
The second topic we tackled in this first year of project is the initial data ingestion and the definition of a basic data format based on parquet format [4] that will serve as a baseline for benchmarking optimization and evolution of the data storage and structure in the course of the project.

As mentioned before, the today available data are around 4000 flights each one with an 10k to 30k parameters sampled at variable sampling rates ranging between 50Hz and 512Hz. Each flight, with an average size of 50-100Gb, is logically and physically divided in “Conditions” each one holding the data of a specific flight operation, condition or manoeuvre that has been performed and properly marked during the recording.

An initial analysis of the source data format and of the queries we expect to execute frequently allowed to define the best target format to use in the ADMITTED cluster thus optimising what was existing in the source database. In particular, the raw data format is more oriented to space saving and optimised to store and retrieve entire flights. On the contrary, ADMITTED algorithms will need to access directly to only few parameters at time but at the same time to be able to process them in memory. We have then defined a new data format that addresses these issues by partitioning data with a finer granularity allowing to search and retrieve with superior performance only the desired parameters belonging to flights and conditions that satisfy certain user-defined target criteria specified at runtime.

This new format, in fact, supports the following desirable features for the typical analysis that would be performed along ADMITTED project:

- It allows to extract and select flights with a subset of parameters without loading everything in memory.
- It allows to store new metadata (e.g., mean, harmonics)
- It allows to cache newly computed parameters (e.g., fft, filtered signals…)
- It allows to update parameters of flight data
- It allows for optimisations for parameters sampled at constant frequency

Finally, since the most typical usage foreseen will exploit just a few parameters, selected flights and conditions data will be loaded into memory using an efficient hierarchy of dict-like data objects.

A third important result is the creation of the Query Catalogue which realize the definition of data containers, classes and algorithms to be used in the implementation of the most popular queries. The need from the business is to structure the access to the same data lake. In fact, several company functions need to access to the flight data and each of them needs a portion of them, integrated with external sources to accomplish different tasks. A building block approach have been defined and structured in this deliverable in order to standardize the data access and query builder by data scientists. The advantage of this approach is to provide a common basis to any company function which need to extract values from data. Data scientist will be able to re-use the building blocks becoming more efficient without re-inventing the wheel, with high efficacy because building blocks will be bug-proof and providing an easy-to-maintain high level algorithm.

An initial set of building blocks has been provided and during the five years of the project there will be a natural evolution of the building blocks. For that the document will be kept open to incorporate all the following building blocks query design and implementation which will be realised in next project months. In order to prove the building block approach a set of use cases has been implemented to evaluate the effectiveness of the usage of building blocks and how they can be combined to implement more complex and insightful algorithms or analysis. The findings show a reduction of development time of 20% about queries development and of 60% in the overall development time from the request and the execution of the first tested query. The high reduction of the overall time is due to the fact that all the connectors and cyber security access have been already validated from low level queries access and quickly replicated in the most complex ones.
4. Conclusions and next steps

The project addressed a highly innovative digital transformation goal in which engineering and test flights are connected thanks to a big data analysis platform allowing to connect flight data and external data sources and extracting knowledge to be used to improve the design of the vehicle. The digital transformation steps achieved during the first year of the project are focused on the settings of the basic infrastructure including the whole cluster both in terms of hardware and software as well as, from the data science point of view, the implementation of a multi-layered building block query approach which allow to different actors access data in a structured and secure way. Complex queries will be so implemented re-using the available APIs of the lower layers. During the next four years of the project the consortium will focus on the implementation of the computation of prediction or suggestion based on data collected and connected and the implementation of novel predictive algorithms based on machine learning techniques.

Acknowledgments
This project ADMITTED has received funding from the Clean Sky 2 Joint Undertaking under the European Union’s Horizon 2020 research and innovation programme under grant agreement No GA832003

References
[1] Clean Sky 2 Information Day dedicated to the 8thCall for Proposal (CFP08), FAST ROTORCRAFT IADPTopics related to FRC WP1 – NextGenerationCivil TiltRotorProject, Leonardo Helicopters
[2] Clean Sky joint undertaking - 8th Call for Proposals (CFP08) - List and Full Description of Topics, April 2018
[3] ADMITTED consortium website: www.admitted-project.eu
[4] Apache Parquet. www.parquet.apache.org