HELICOPTER BIG DATA PROCESSING AND PREDICTIVE ANALYTICS: FEEDBACK & PERSPECTIVES

Ammar Mechouche, Nassia Daouayry, Valerio Camerini
Ammar.Mechouche@Airbus.com
Airbus Helicopters
International Airport of Marseille Provence
Marignane (France)

Abstract
This paper offers a comprehensive return of experience on the deployment of big data technologies enabling various descriptive and predictive analytics within the helicopter industry. It shows how these technologies can efficiently be employed to allow storing and processing the large quantity of flight data which are made available, and consequently how they offer additional analytics capabilities to the analysts. In order to demonstrate these benefits, the paper presents applications concerning statistical fleet data analysis and predictive maintenance. Concluding remarks are then given with particular attention to limitations of distributed big data technologies and challenges regarding their adoption within the industry.

1. ABBREVIATIONS

AH – Airbus Helicopters
TBO – Time Between Overhaul
NFS – Network File System
HDFS – Hadoop Distributed File System
CPU – Central Processing Unit
RDBMS – Relational Database Management System
I/O – Input / Output
ETL – Extract Transform Load
DAG – Directed Acyclic Graph
SQL – Structured Query Language
FDCR – Flight Data Continuous Recorder
HUMS – Health and Usage Monitoring System
MIS – Maintenance Information System
MRO – Maintenance Repair and Overhaul
MGB – Main Gear Box

2. INTRODUCTION

The subjects of big data and predictive analytics are recently attracting significant attention from the industry [1]. Large investments are made in research activities revolving around these topics by companies, start-ups, industrial groups and governments. Efforts are made in the development of methodologies, systems and tools enabling for collecting and efficiently analyzing a massive amount of heterogeneous data from different sources, with the aim of extracting valuable, non-obvious information which might be used for improving products, offering new services or gaining additional business insights. Within the helicopter industry, a huge volume of customer data is collected from the Health and Usage Monitoring Systems (HUMS) and from the Flight Data Continuous Recorder (FDCR). Specifically, hundreds of parameters are recorded from the FDCR, coming from each of the vehicle systems connected to the avionics, whereas HUMS data mostly consist of counters monitoring quantities related to the usage of the machine, and high-frequency vibration measurements acquired from multiple sensors, typically located in the proximity of mechanical transmission components and dynamic systems. Airbus Helicopters gathers data from hundreds of connected helicopters, operated by several customers worldwide and performing a large variety of missions. The analysis of such data brings considerable benefits in terms e.g. of safety and operations enhancement, maintenance optimization, system design improvement and in-service incidents support. The objective of this paper is to investigate how big data technologies can efficiently be exploited to allow storing and processing the large quantity of data which are made available, with a focus on FDCR data. Issues concerning data volume, variety and generation velocity are outlined, showing the insufficiency of classical data storage and processing systems and pointing to the need for big data infrastructures and tools. The adopted big data technologies are then discussed, with a focus on the computational

Presented at 45th European Rotorcraft Forum, Warsaw, Poland, 17-20 September, 2019

benefits and the additional processing capabilities which are offered to the analysts. Specifically, the implementation of tools and architectures based on Spark, Spark ML / ML Lib, Map-Reduce, Tez and Hadoop within Airbus Helicopters is discussed. Such possibilities are exploited to speed-up large scale statistical analysis and the development / deployment of predictive analytics algorithms, able of performing regression and classification tasks over all the historical fleet data. Computational advantages of big data technologies are thoroughly discussed, along with the drawbacks associated to the costs of system maintenance and development. Based on such technologies, automatic algorithms leveraging on the combination of expert domain knowledge and data-driven methods are deployed on the huge volume of available data, leading to the achievement of significant benefits for safety and maintenance operations. Applications concerning statistical fleet data analysis and predictive maintenance are presented. First, the capabilities of the developed processing system are demonstrated on a case involving statistical analysis of a challenging volume of fleet data. Flights from all historical data for a helicopter type are analyzed in order to accurately quantify the actual percentage of time spent in a specific regime, allowing the validation of a design hypothesis. Then, a case for rotor brake maintenance anticipation is investigated. It is shown how the big data technology allowed for verifying an expert hypothesis on the full historical data, reliably setting up a predictive maintenance indicator. Finally, some research activities related to the MGB normal behaviour tracking and virtual sensing application are explored, where multi-parameter time-series regression is performed with the aim of detecting lubrication system problems. Specifically, a case of oil contamination detection is showed. As a result, this paper offers a comprehensive return of experience on the deployment of big data technologies enabling various descriptive and predictive analytics within the helicopter industry. Concluding remarks are then given with particular attention to limitations of distributed big data technologies and challenges regarding their adoption within the industry.

3. THEORETICAL BACKGROUND

This section provides an overview of AH collected data. It highlights the need for big data technologies in order to better manage and analyse this data. Then, it gives a brief state of the art of big data technologies that were tested to speed-up AH helicopter data processing.

3.1 Problem definition

Figure 1 gives a brief overview of the collected HUMS data at Airbus Helicopters, highlighting the high volume generated over a wide range of machines flown by multiple operators for different missions. As we can see, FDCR data represent more than four hundred thousand flight hours. This data was managed only using a RDBMS technology. In fact, the HUMS application has two main components, namely decoding tool and computing modules. The decoding tool is responsible for extracting engineering values from the raw data and stores them in a SQL database table. The computing modules are Java applications, SQL procedures or Python scripts. They compute results which are stored in a database as SQL tables. These results are then retrieved by the web applications and presented to the requesting users. Due to the limitations of the classical technologies, engineering data - which consist of data useful for engineering analyses - is systematically removed when computing modules have finished calculating the results. Thus, when a new computing module is deployed and/or a computing module is modified, engineering values must be extracted again as they are not stored in the database. As a consequence, computational time is increased, leading to reduced analysis efficiency. Moreover, as the SQL data infrastructure is customer oriented, big data analytics are very limited in order to keep it efficient and highly available. For all these reasons, the deployment of a big data platform was necessary.

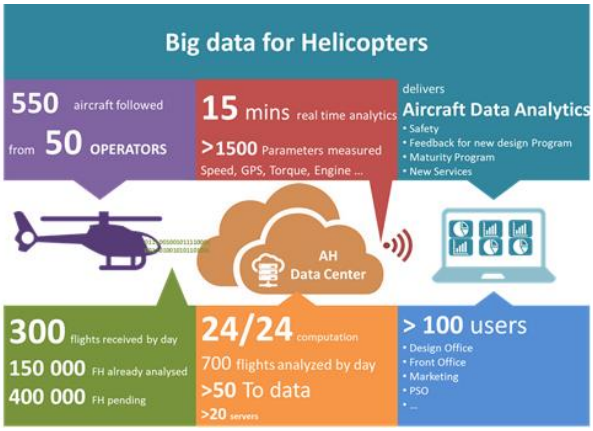

Figure 1: HUMS flight data statistics

3.2 Big Data Technologies

Big Data technologies are the result of many efforts done by the research community and digital companies, in order to deal with the growing amount of data generated by 'democratized' connected objects, social networks, industries, public administrations, etc. With classical technologies, e.g. relational database management systems (RDBMS) based storage, the data are moved from storage to the application server in order to be processed, which makes it not scalable. Indeed, traditional technologies have limited storage capacity, rigid management tools and are expensive [2]. On the contrary, big data technologies can support distributed processing across large datasets. They also support the creation of scalable storage. Hadoop[1] for example, which is the de-facto standard for most of the big data based solution across industries, provides a distributed storage system that supports storing large amount of data in multiple nodes. Large datasets are split into 'blocks' and stored across multiple machines. Blocks are then processed by independent, idempotent processes (Maps) in parallel. Hadoop Map-Reduce (MR) programming paradigm automates the process of moving the computation logic to multiple nodes and executes them. So, it can move the processing to the node where required data is already present (Data locality awareness) [4]. In the first generation Hadoop system (called Hadoopv1), the resource management component is tightly coupled with the programming model. However, in the second generation Hadoop system (called Hadoopv2), the resource manager is decoupled from the programming model so that the resource manager can provide its service to different programming models. The resource manager in Hadoopv2 is called YARN.

Several projects that use Hadoop Distributed File system (HDFS) and its programming model are developed. Some of them are listed below:

- HBase: A scalable, distributed database that supports structured data storage for large table.
- Hive: Data warehouse infrastructure that provides data summarization and ad hoc querying.
- Pig: A high-level data-flow language and execution framework for parallel computation.
- Ambari: A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop Map-Reduce, Hive, HBase, Oozie[2], Pig[3], etc. Ambari also provides a a user-friendly dashboard for viewing cluster health such as heatmaps and ability to view Map-Reduce, Pig and Hive applications visually along with features to diagnose their performance characteristics.
- Tez: A generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute an arbitrary Direct Acyclic Graph (DAG) of tasks to process data for both batch and interactive use cases.

Another important big data technology standard is Apache Spark engine. It is a fast and general purpose processing engine for large scale data processing. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, Machine Learning, stream processing, and graph computation. Spark programming model is seen as an alternative to the Map-Reduce programming model used in Hadoop. Map-Reduce programming model is best suited for batch processing application. Hence, many real time applications and applications that involve iterations of map and reduce operations (Ex: Machine Learning applications) could not be efficiently migrated on a big data infrastructure. Hence, a different and faster programming model was needed. Apache Spark was developed from the scratch as a general purpose processing engine. It has an advanced DAG execution engine that supports cyclic data flow and in-memory computing. It runs on Hadoop, in a standalone mode or in the cloud as well. It can also access diverse data sources including HDFS, HBase, S3, etc. A detailed review on big data technologies, including several related challenges and opportunities, can be found in the references [2, 3, 5].

4. AIRBUS HELICOPTERS SOLUTION

The development of the big data solution was carried out in two steps. First, a demonstrator was developed using a laboratory big data cluster (composed of 6 nodes) and a subset of FDCR data. Different architectures were benchmarked and tested [4]. The retained solution consists of Apache Spark for distributing the process of extracting engineering values, Apache Hive for accomplishing SQL style database warehousing activities, and Apache Hadoop for scalable and distributed storage mechanism.
The main motivation behind the selection of Apache Hive was the fact that the existing HUMS application involves many RDBMS based data warehousing activities. Thus, to perform similar activities at a larger scale and to seamlessly migrate towards big data platform, Hive was selected for its similarity with RDBMS for what concerns the query language. A part of the existing computing modules was able to translate into

[1] https://hadoop.apache.org/

[2] https://oozie.apache.org/

[3] https://pig.apache.org/

HiveQL modules and successfully run in the big data platform. With Apache Hive, the learning gap, that is migrating from SQL to HiveQL, is very short and can encourage the engineers to adopt big data solution without spending much time. Moreover, Apache Hive is more suitable (than HBase for example) to perform aggregation operations and scanning of records, which represent an important part of the expected flight data analysis, as it will be illustrated in the next section.

5. RESULTS

5.1 Big data benchmarking

For performance benchmarking, a SQL based infrastructure was re-created and successfully tested on a material that has similar characteristics as the cluster nodes. This makes it easier to compare the existing design and the tested one. Then, performance benchmarking of Hive based computing modules was done on all the different execution engine configurations of Hive. Figure 2 shows the obtained results recorded on various scenarios.

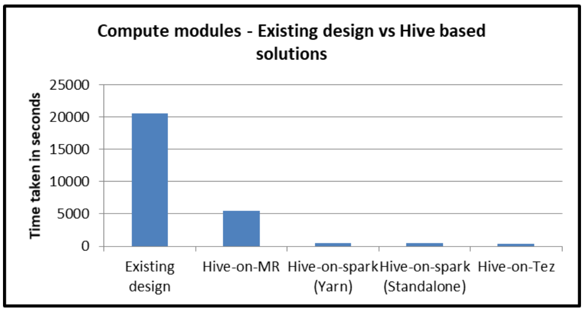

Figure 2: Performance benchmarking of various Hive execution engine configurations [4].

Hive on Spark and Hive on Tez got the best results and reduced processing time by a factor of ~54. Spark being the standard and more adopted in the big data community, it was selected for *the* deployed solution. Performance benchmarking was also done for the decoding phase, and similar results were obtained. Here, we focus on computing module step because in the big data solution all data is decoded once and made available online for analysis, whereas in the existing solution decoding should be performed for every created / modified computing module.

As a second step, the retained architecture was deployed within the AIRBUS "Skywise[4] on premise" big data platform, which contains much more nodes (several tens of nodes) than the laboratory cluster. As a result, the processing time benefits are more important.

Actually, both of the platforms (RDBMS and big data) are synchronized and live together as shown on Figure 3. The RDMS platform is customer-oriented, it is optimized to quickly processing new coming flights data and to make the results available in the web applications. The big data platform is mainly used for research activities, algorithms / indicators development and maturation, hypothesis testing, large scale statistical analysis. Once an algorithm or a maintenance indicator [5] is matured on the big data platform it is then implemented in the RDMS platform in order to make the results - continuously for new flights - available through the web applications. This cohabitation of the platforms is necessary because the redevelopment of the existing web applications on the big data platform is costly and would take a long time.

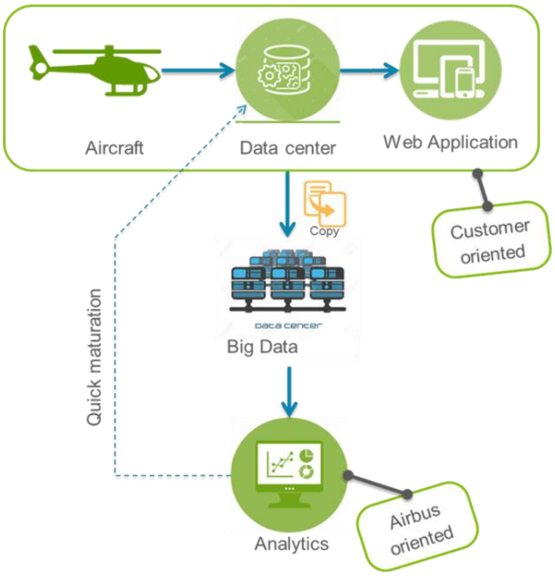

Figure 3: RDBMS and big data platforms

5.2 Benefits for the analytics

Predictive analytics encompasses a variety of statistical techniques from data mining, predictive modelling, and Machine Learning that analyze current and historical facts to make predictions about future or otherwise unknown events[6]. This section provides a feedback within the helicopter industry through a variety of examples that

[4] https://www.airbus.com/aircraft/support-services/skywise.html

[5] Indicators that allow anticipation of failures.

[6] https://en.wikipedia.org/wiki/Predictive_analytics

demonstrate the added value of big data technologies for predictive analytics.

5.2.1 Whole Historical Data Replay

The immediate benefit from the implementation of the big data platform was the ability for the computing modules to be executed on the whole fleet data with only minor efforts, and quickly make the results available for users. In fact, writing a script and computing results is now a matter of minutes, or hours in worst cases. For complex computing modules (such as Java / Python applications) that continue to be developed in a classical way, their encapsulation within the big data infrastructure is quite easy. In fact, instead of reading from NFS file system, minor modifications allow them to read from HDFS and then to be parallelized. Thus, results on the whole fleet data can be obtained very quickly for every Java computing module.

As an example, at Airbus Helicopters engineers develop digital twins, some of which consist in calculating the wear of some mechanical parts at each flight. Then, an alert is raised if the cumulative wear over time exceeds a predefined threshold. In this case, when such a digital twin is deployed it is important to compute values for the entire historical data in order to determine the cumulative values since the installation of the monitored mechanical parts (Figure 4). Otherwise, the deployed digital twin could miss alerting for all monitored parts currently flying.

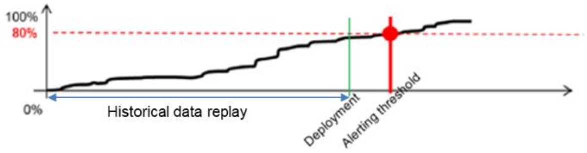

Figure 4: Example of cumulative indicator

5.2.2 Statistical Analysis at Fleet Level

This section shows an example of statistics computed on all historical FDCR data through the big data processing capabilities. Figure 5 shows the actual percentage of time spent in a specific regime for all the machines of a selected fleet. Such a flight spectrum allowed the validation of a design hypothesis for all flights already performed by in-service helicopters. This statistics would typically take several months of deployment and processing time to be computed using traditional architecture, while it can be obtained within a few minutes using the big data platform.

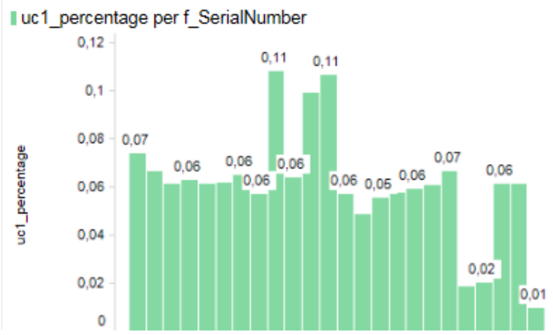

Figure 5: Statistics on whole FDCR fleet data

5.2.3 Maintenance Anticipation Hypothesis Testing

This section shows how big data technologies allow the rapid development of indicators for failure anticipation and predictive maintenance, thanks to fast testing of expert insights. To illustrate this, an example related to rotor brake adjustment is here considered. In fact, for some flying rotor brake systems, an adjustment is required periodically after a certain wear of the rotor brake pads until their removal. This adjustment is done from customer side based on the pilot assessment about the rotor brake response. This information is tagged within the Maintenance Information System (MIS) as: Rotor brake weak, rotor brake slow to respond, rotor brake too efficient etc. Then, the idea of Airbus Helicopters engineers was to monitor the braking duration in order to support customers to optimize, anticipate and plan adjustments actions. Indeed, the objective is twofold: first, to inform customers when it is time to readjust the rotor brake, and second, to ensure that the performed adjustment is not hard, eventually increasing the risk of rotor brake pad crack. The implementation of the algorithm computing the rotor brake duration from flight data was very simple, and the big data platform allowed testing and validating engineers' hypothesis by computing – in few minutes only – the brake duration for the whole historical flights (Figure 6). During the test phase, the deployed indicator was able to anticipate more than 85% rotor brake adjustments several days before the adjustment date logged into the customers Maintenance Information System (MIS).

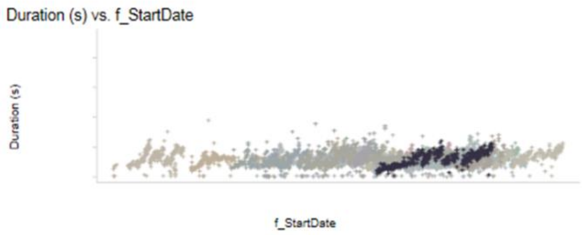

Figure 6: Rotor brake maintenance anticipation

5.2.4 MGB Operation Normality Monitoring

From research point of view, the big data platform opened new perspectives. The lead-time to test new ideas is significantly reduced, leading to the acceleration of indicators development related to helicopters predictive maintenance. In [6], authors presented a methodology to monitor the normal behavior of MGB oil and pressure parameters. The methodology consists of visualizing concerned and contextualized parameters using a sliding window and co-occurrence matrices (Figure 7). The assumption was that in similar operation conditions the correlation of the normalized MGB oil and temperature will be preserved over time. Then, when this correlation is no longer valid, abnormality within MGB lubrication system is suspected. The development of this methodology required to have FDCR data online as well as a high processing performance, that were brought by the big data platform.

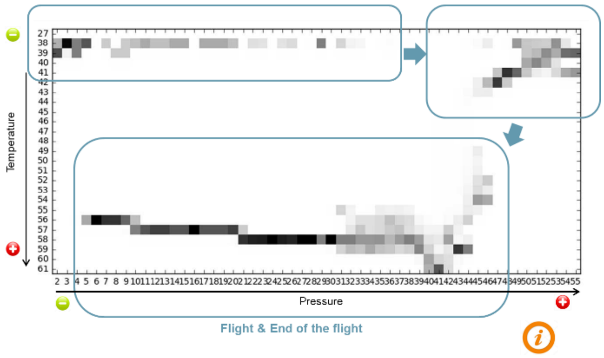

Figure 7: Co-occurrence matrix based abnormal bahavior detection

5.2.5 **Virtual Sensing Exploration**

This section presents preliminary results obtained with Spark ML / ML Lib distributed Machine Learning environment on a multi-parameter time-series regression analysis case, without any data reduction operation (Figure 8). The analysis is performed with the aim of detecting lubrication system problems.

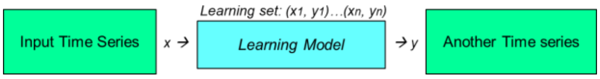

Figure 8: overall learning process

This ongoing work consists of the following steps:

- Determining parameters that influence the response parameter, using the entire available data;
- Learning a model that predicts the response parameter, using entire fleet time series data that do not contain anomalies regarding lubrication system, thanks to a distributed learning allowed by Spark ML / ML Lib;
- Validating the predictive model;
- Defining an alerting threshold;
- Deploying the model for real-time fleet analysis.

Results obtained with a preliminary version of the model were promising. In fact, when tested on a known anomaly related to the MGB oil contamination, the discrepancy between the results predicted by the model and the data collected from the sensors enabled to highlight and detect the anomalous behaviour as an offset. Figure 9 shows data from two situations: 1) a “normal” case where the measured and predicted times series, respectively black and blue, are very similar; 2) an “anomalous” case presenting oil contamination, where a significant difference exists between the ranges of measured and predicted time series, respectively red and green.

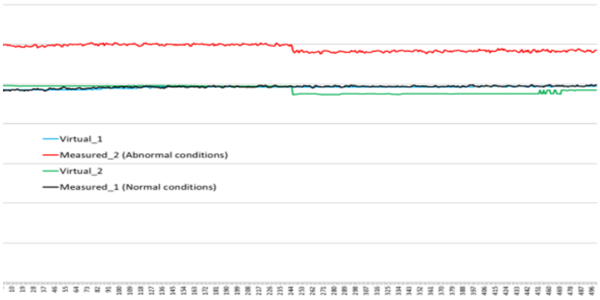

Figure 9: Example of prediction carried-out by the model

6. DISCUSSION

The computational benefits offered by the application of big data technologies to helicopter predictive analytics have been explored through this experience. It was shown that the significant improvement over classical computing architectures could directly result into new analysis capabilities. The faster return of experience that can be provided to the experts by analysing the totality of fleet data leads to a significant speed-up of troubleshooting, incident investigations and technical solutions development. This proves the big data tools as an important asset for the analysts within the helicopter industry.

However, as explained in [2], CPU and disk drives performances are doubling each 18 months while the I/O operations do not follow the same pattern. This may slow accessing data and more generally affect the performance and scalability of big data applications. Thus, intelligent data management and use of these technologies will be required. A solution could be the collection of only data that is of interest for the targeted domain. The difficulty in this case is to define and identify proactively what is the relevant data which will be needed in the future. In the helicopter industry, this could be reflected by a close collaboration with system design responsible when deciding about data to collect. Another solution could be the collection of all possible data, and incrementally filter the processing only to the sub-part required in future analysis. This latter solution could however slow the data analysis process, since it adds a recurrent data management phase (decoding, storing, configuration, etc.).

Libraries like Spark ML bring new capabilities for doing distributed Machine Learning at large scale. However, some Machine Learning algorithms are still not included in these libraries due to the theoretical challenges for them to be parallelized (such as deep learning algorithms). This limitation, from the experience point of view reported here, should not reduce the interest of these libraries and the engagement of data scientists towards them. In fact, the most important regression and classification algorithms are already managed[7]. For the remaining ones, they can be used through classical libraries such as Scikit-Learn[8] using small data and take benefits from big data technologies for testing and validating the learned models on big amounts of data.

Finally, another challenging topic is the efficient setting, maintenance and adaptation of on premise big data technologies, considering this very fast evolving domain and the variety of developed technologies. It represents a high investment for enterprises. An interesting option is to move to the more and more growing cloud-based analytics and cloud computing. However, despite the efforts done by the big data community to mitigate risk of cloud security violations risk (regarding data access, encryption, computation verification, etc.), this still is considered as a serious threat for many enterprises that have sensitive data but could not invest for their own big data infrastructures, and make them hesitating to adopt these technologies.

7. CONCLUSION

This paper offers a comprehensive feedback on the deployment of big data technologies enabling various descriptive and predictive analytics within the helicopter industry. Various examples, ranging from simple statistics to sophisticated Machine Learning algorithms, have demonstrated the added value of these technologies. The implemented big data platform allows now developing more Machine Learning and intelligent algorithms that are able for extracting more valuable Knowledge from massive time series data collected from AH in-service helicopters.

Acknowledgement: We would like to thank our colleagues Balachandar AMARNATH, Pierre-Loic MAISONNEUVE and Valentin AUBIN for their active role in the development and deployment of the Airbus Helicopters Big Flight Data solution.

8. REFERENCES

[1] Lee Jay, Kao Hung-An, Yang Shanhu. 2014. Service Innovation and Smart Analytics for Industry 4.0 and Big Data Environment. 16. 3–8.

[2] Oussous Ahmed, Benjelloun Fatima-Zahra, Ait Lahcen Ayoub, Belfkih Samir. 2017. Big Data Technologies: A Survey. Journal of King Saud University - Computer and Information Sciences.

[3] Alexandros Labrinidis and H. V. Jagadish. 2012. Challenges and opportunities with big data. Proc. VLDB Endow. 5, 12, 2032-2033.

[4] Balachandar AMARNATH. 2016. Big data based infrastructure management for HUMS use case of Airbus Helicopters. AIX/2016-09/001. Internal AIRBUS technical Report.

[5] Ahmad Ghazal, Tilmann Rabl, Minqing Hu, Francois Raab, Meikel Poess, Alain Crolotte, and Hans-Arno Jacobsen. 2013. BigBench: towards an industry standard benchmark for big data analytics. In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD '13).

[6] Nassia Daouayry, Pierre-Loic Maisonneuve, Ammar Mechouche, Vasile-Marian Scuturici, Jean-Marc Petit. 2018. Predictive Maintenance for Helicopter from Usage Data: Application to Main Gear Box - European Rotorcraft Forum (ERF). Delft.

[7] https://spark.apache.org/docs/2.2.0/ml-classification-regression.html

[8] https://scikit-learn.org/