Paper 1049-2017

Exploiting Competitor Data Using SAS/ACCESS® Interface to Hadoop

Kayne Putman, British Airways

ABSTRACT

The British Airways (BA) revenue management team is responsible for surfacing prices made available in the market with the objective of maximizing revenue. BA is currently working to understand how publicly available competitor data can help facilitate better decision making. Due to the low level of aggregation, competitor data is too large (and consequently too expensive) to store on conventional relational databases. Therefore, it has been stored on a small Hadoop installation at BA. Thanks to SAS/ACCESS® Interface to Hadoop, we have been able to run our complex algorithms on these large data sets without changing the way we work and whilst exploiting the full capabilities of SAS®.

INTRODUCTION

BA is one of the world's leading airlines with billions of pounds in annual revenue from transporting over forty million passengers from their hubs in London, UK. Operational Research (OR) is an internal consultancy who provide BA with analytical excellence, specialising in problem structuring, business modelling and complex data analysis. Our group's mission is to guide effective decision making, drive delivery of the business plan and to optimise KPIs.

In this paper, we share a success story of how SAS software was used to interrogate data which has helped guide better revenue management decisions. Specifically, we will discuss use of SAS/ACCESS Interface to Hadoop.

This paper is intended to demonstrate:

1. The ease of use of use SAS® as an analytics platform to interact with Hadoop
2. Demonstrating Hadoop as a data value staging area
3. Convey real business benefit of Big Data

Please note that any datasets or plots portrayed below are purely for illustration purposes, whilst the business problem at hand was real and results are in line with those presented to the business.

REVENUE MANAGEMENT AT BRITISH AIRWAYS

Whilst many airlines opt for "off-the-shelf" revenue management systems, British Airways has built its own system from the ground up. Over the last sixty years, Operational Research have been critical in maintaining & continually improving this system which contains a host of varied forecasting and optimisation algorithms to govern how inventory and pricing decisions are made. By hosting this solution internally, we are able to be more dynamic in exploring the value of changes to either the algorithms or the underlying data which is employed.

Simply put, the task of revenue management is to maximise total revenue generated for each flight. There are two levers available to manage this:

1. How many seats are made available at each point in time before a flight will depart
2. The price available at each point in time before a flight will depart.

Nevertheless, the two levers are not entirely independent. There are a number of other KPIs which revenue management use to track the effectiveness of a change. These include:

1. Yield – the average price achieved per sector, e.g. if 200 passengers on a plane generate £100,000, then the yield is £100,000 / 200 = £500.

2. Revenue per available seat kilometre (RASK) – a unitized metric to adjust for different flight distances, e.g. if the above was on an aircraft with 250 seats to New York (which is 5,554 Great Circle kilometres from London), then the RASK would be £100,000 / (250 * 5,554) = £0.072.

Due to the rich amount of data which can be theoretically employed in revenue management, it makes it a great candidate for Big Data investigations as we are consistently challenging ourselves to find the optimal level of detail of data to guide everyday revenue management decisions.

THE RISE AND ADOPTION OF BIG DATA SOLUTIONS AT BRITISH AIRWAYS

Big data is a phenomenon which has been recently omitted from the Gartner Hype Cycle for Emerging Technologies[1] as businesses are starting to demonstrate sizeable benefits from it; and British Airways has certainly observing the value of it within our business.

OVERVIEW OF ANALYTICS ARCHITECTURE AT BRITISH AIRWAYS

For over 20 years, British Airways have had large Teradata and Oracle systems to store all business data.

In 2014, British Airways invested in a small Hadoop installation with a HortonWorks® distribution of nine nodes. The initial business case (for the investment) was made by our Business Intelligence (BI) team and was predicated on more efficient storage of archived data which needs to be infrequently queried. Thus, we have adopted a strategy of Hadoop co-existing with our RDBMS.

BA has had a relationship with SAS for over 20 years and at the time of writing, we have a parallel 9.3 & 9.4 installation, with intentions of migrating into one version in Q1 2017. SAS is one of a couple of analytics platforms which enable analysts to access, manipulate & model data through these three data warehouses. Nevertheless, with a number of internal success stories in purely deploying open-source analytics solutions, BA is currently reviewing a tool strategy increasingly focused on open-source architecture.

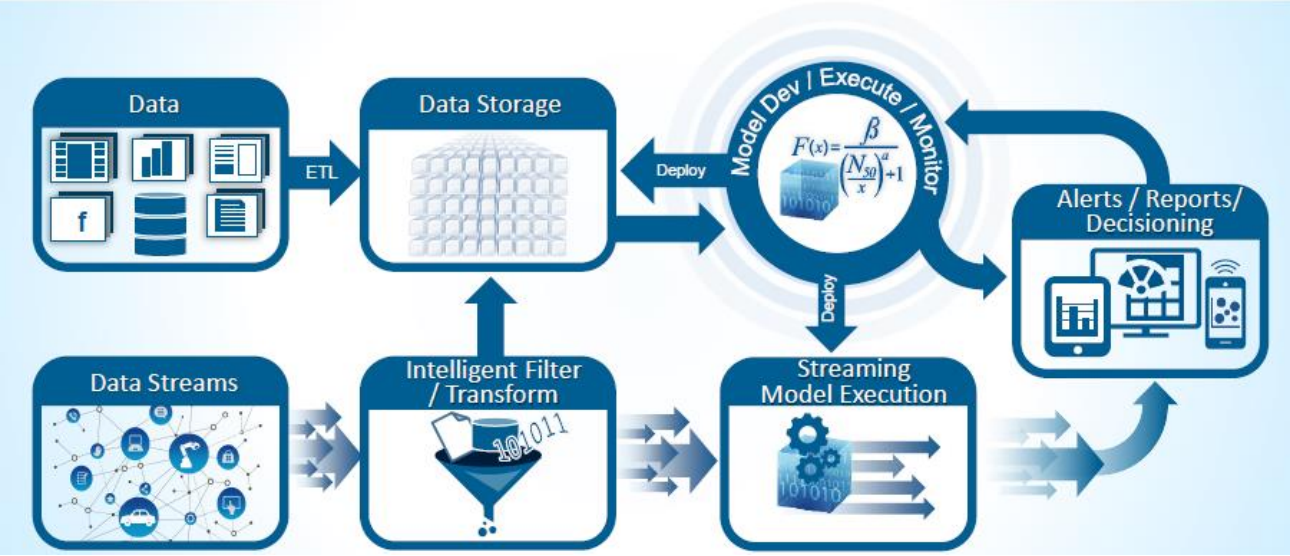

Figure 1. Analytics Architecture at British Airways (graphic courtesy of SAS)

FACILITATING & FOSTERING INNOVATION

Since the introduction of the wider Hadoop ecosystem, OR & BI have been instrumental in proactively identifying opportunities which can enable us to work more effectively or efficiently or drive incremental value to our clients specifically across the commercial, customer and operational directorates.

In 2016, over seventy business use cases were identified through workshops where reduced data latency (velocity), removed data size blockers (volume) & varied unstructured data (variety) can help British Airways

use new data and interrogate our data in a more intelligent manner using our Hadoop instance as a staging area to determine the value of these data sources. This test-and-learn philosophy has enabled OR and BI to be more innovative and to move faster as an analytical unit.

Figure 2. Embedding a new way of working at BA

AN EVOLVING SKILLSET FOR ANALYSTS

For many, Hadoop offers a welcomed solution, allowing businesses to store enormous amounts of data on clusters of commodity hardware. However, whilst this open-source software has become widely popular for housing big data, connecting that data to analytics and decision-making poses a challenge.

Naturally, a lot of analysts feared they would have to learn a new technology or language. However, thanks to the clever folks at Facebook, and subsequently the Apache Foundation, a new language coined Apache Hive (often referred to as HiveQL) has enabled analysts to extract data from HDFS in an SQL-type language, thus removing the barrier to entry of the Hadoop technology.

Whilst Hive subtly differs to other SQL languages, the syntax is more-or-less the same, and negates the need to navigate MapReduce or Pig (other ETL languages) for simple extraction & manipulation of relational data.

Nevertheless, at some stage, a more evolved “super-user” will be required to perform tasks over-and-above the simple extraction & extraction of data from Hive tables. For example: manipulating unstructured data, creating schemas and structuring data into Hive tables from unstructured sources & embedding real-time analytics and decision-making where the data resides. This is where other Hadoop libraries such as Flume, Spark & Storm start to come to life.

In our experience, only one in ten analysts who are proficient in Hive will need to know these more evolved functions, although it is anticipated this may change going forward.

OUR HADOOP USE CASE: USING COMPETITOR DATA

To manage our flights we use a range of datasets with varying levels of granularity, for both systems and decision support. We want to determine what level of detail was most useful to make everyday pricing decisions. We explored a range of rich sample data sources to test this.

In the past, data size, cost-consciousness & long BI project lifecycles made it hard to explore the value of additional, very detailed data sources in an agile way. However, since two of those blockers (data size & long BI lead-times) could now be unblocked by using Hadoop it was now possible to understand on a trial basis the value of that data as part of the ongoing revenue management decision making process. It was agreed that OR would timebox a fixed period of time spend interrogating the data and determining whether it would be worth procuring in the long-term.

Below contains further detail how the data was processed, extracted and modelled to derive a conclusion on the benefit of the data.

HOW WAS THE DATA PIPED AND MANAGED

Hourly compressed files (gzip) were transferred via SFTP from an InFare server to a "utility" server within the Hadoop cluster on the BA internal network. From here, the files were moved directly into HDFS under a folder for that day. HDFS was also synchronised with AWS s3 as an archive.

Using Hive SQL, three external tables were then created giving various different partitioned views of the data. A file format specific SerDe was used to automatically remove the header from the file and use a non-standard column separator so there was no need for an ETL process for the data before it was available to analysts.

In total, over two years' worth of data was obtained, sizing to over 4,000,000,000 rows and 82GB (compressed).

HOW WAS THE DATA ACCESSED

Through SAS 9.4 and SAS/ACCESS Interface to Hadoop, we were able to extract data from HDFS. This can be achieved by using implicit or explicit pass-through.

Implicit pass-through refers to where a library is created and the PROC or DATA steps are translated (where possible into the native SQL), like below:

```
LIBNAME HDP2 HADOOP SERVER="HDP_SERVER" USER=XXX PASSWORD=XXX PORT=XXX
SCHEMA=XXX SUBPROTOCOL=HIVE2;
LIBNAME HDP2 REMOTE SERVER = UNIXID;

PROC PRINT DATA = HDP2.OLD_DATA;
PROC SQL; CREATE TABLE NEW_DATA AS SELECT * FROM HDP2.NEW_DATA;
```

Conversely, explicit pass-through is where a wrapper facilitates coding in the native language of the database, which is executed exactly as coded – therefore is based on the premise that SAS is not used to alter or translate the code. An example of where this is done is below:

```
PROC SQL;
CONNECT TO HADOOP (SERVER="HDP_SERVER" AUTHDOMAIN=HDP_LIVE PORT=XXX
SCHEMA=XXX SUBPROTOCOL=HIVE2);
CREATE TABLE OLD_DATA AS
SELECT * FROM CONNECTION TO HADOOP
(SELECT * FROM SCHEMA.OLD_DATA);
DISCONNECT FROM HADOOP;
QUIT;
```

For further details about explicit and implicit pass-through, please see this paper[2].

The majority of querying was performed using explicit pass-through, as we wanted to upskill in the difference between Hive SQL and Teradata SQL (which is predominantly what we use). We found a very useful guide to help us with this from HortonWorks[3]. We found that we could mirror the majority of commands and functions available in Teradata. and we were satisfied with the amount of processing which could be performed where the data resides, thus only bringing back the necessary information to the SAS libraries on the Unix server.

Function	Teradata	Oracle	Hive
Row limits	SELECT TOP 10 * FROM table ORDER BY col1	SELECT * FROM (SELECT * FROM table ORDER BY col1) WHERE ROWNUM <= 10	SELECT * FROM table ORDER BY col1 LIMIT 10
Sampling	SELECT * FROM table SAMPLE 0.05	SELECT * FROM table SAMPLE (5)	SELECT * FROM table TABLESAMPLE (5 PERCENT)
Date and Timestamp calculations	WHERE datecol = DATE '2016-01-01'	WHERE datecol = DATE '2016-01-01' WHERE datecol = TO_DATE('2016-01-01','yyyy-mm-dd')	WHERE FROM_UNIXTIME(UNIX_TIMESTAMP(datecol), 'yyyy-MM-dd') = '2016-01-01'

Table 1. Some tips on the difference between Teradata, Oracle & Hive SQL

HOW WAS THE DATA MODELLED & ACTIONED?

Ultimately, the objective of this investigation was to identify if competitor price data improved our decision making in Revenue Management.

The new information was manipulated and modelled in various ways to build a dataset which provided views of data to the route controllers in revenue management which they had not previously seen. This enabled them to make more informed decisions and hypothetically lead to improved revenue decisions on their routes.

We sought feedback from the people involved and also measured whether revenue on their routes improved to decide the shape of data and decision support required going forward.

CONCLUSION

Through this paper, we have shared how SAS® software has been used to unlock more value from our Hadoop installation to exploit novel data to optimise revenue management decision making.

This choice of technology reduced was beneficial two-fold:

1. required level of BI intervention, thus timelines
2. enabled analysts to interrogate the data using a familiar analytics platform, thus negating the need for upskilling in a new technology

By performing a number of trials, we were enabled to quickly demonstrate whether additional data improved our decision making to optimise revenue generation.

REFERENCES

[1] dataanami. 2015. “Why Gartner Dropped Big Data Off the Hype Curve”. Accessed Oct 10, 2016. https://www.datanami.com/2015/08/26/why-gartner-dropped-big-data-off-the-hype-curve

[2] Capobianco, F. 2011. “Explicit SQL Pass-Through: Is It Still Useful?” *Proceedings of the SAS Global Forum 2011*, Las Vegas, NV: SAS Global Forum. Available at: http://support.sas.com/resources/papers/proceedings11/105-2011.pdf

[3] HortonWorks. 2016. “Cheat Sheet Hive for SQL Users”. Accessed June 30, 2016. http://hortonworks.com/wp-content/uploads/2016/05/Hortonworks.CheatSheet.SQLtoHive.pdf

ACKNOWLEDGMENTS

We would like to extend our gratification to all clients in British Airways who have allowed this success story to be shared with the a wider SAS audience.

RECOMMENDED READING

- Bodea, T. Ferguson, M. 2014. *Segmentation, Revenue Management and Pricing Analytics*. 1st ed. New York, NY: Taylor & Francis.
- SAS Institute White Paper. Bringing the Power of SAS® to Hadoop. 2014. http://www.sas.com/en_us/whitepapers/bringing-power-of-sas-to-hadoop-105776/download.html#formsuccess
- tutorialspoint. 2016. “Hive – Introduction”. Accessed March 01, 2016. http://www.tutorialspoint.com/hive/hive_introduction.htm

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Kayne Putman
British Airways
kayne.putman@ba.com