Improving Transport Management with Big Data Analytics

* Albert Nagy, ** József Tick
* Applied Informatics, Óbuda University, Budapest, Hungary
** Applied Informatics, Rector's Cabinet, Óbuda University, Budapest, Hungary
albert.nagy@me.com, tick@uni-obuda.hu

***Abstract*— Transport corporations today are rich in data but poor in information. Lack of real business intelligence, may result in lower efficiency, decreased level of travel experience. This paper briefly outlines the course of developments referring to the transport management using big data analytics and their impact on cost effectiveness.**

I. INTRODUCTION

Most of the transport companies have been using ERP and/or other proprietary systems for several decades in order to manage their core business activities. These systems have been producing a huge amount of information and that volume is growing exponentially. Beyond the internal enterprise data, today, even more can be derived from the private or hybrid cloud applications. Another large area is the data in legacy systems as well. Nowadays, big data challenges the norms and culture established in these companies. The information available today has reached the size and complexity that make barriers towards innovation. As computing resources have evolved, advancing to handle data size and complexity better, companies exploit greater benefits from information and analytics. But with the current methods, the conclusions drawn from existing sources, in many cases, result in low efficiency.

Companies should take advantage of new technologies and methods by using advanced predictive models. Connecting data sets in a deeper and complex way through the analysis and deep learning, several areas can be identified to develop new approaches, higher efficiency as well as cost savings.

Huge volumes of data from which useful information can be derived is available – mainly in structured but also in unstructured formats. The available data contains hidden and complex relationships as well as dependencies.

Predictive Analysis enables companies to build predictive models to discover hidden insights and relationships, from which predictions could be made about future events. Special tools and technologies are available to analyze this data and present the results.

This paper aims to give an example for the use of Big Data Analytics through a case study in which analysis have been carried out at the Budapest Transport Privately Held Corporation (BKV). As a result of the examination the directions of the research is outlined based on the observation and experience, aiming more efficient, higher level of services while keeping operating costs low.

II. STORING BIG DATA

Storing large quantity of data efficiently, simple data storage solutions cannot be implemented. Not only because of the amount of data, but also their complexity requires a higher-level concept. Below it is presented an overview of the datasets characteristics and the modern database technologies through vendor analysis.

A. Datasets

The data must be perceived valuable and benefit to the enterprise using it. The data should satisfy the combination of the following 3V characteristics:

Volume: Where the quantity of data to be stored and analyzed is large enough to require special considerations

Variety: It consists of multiple types of data from multiple sources. Structured data held in tables or objects or unstructured data, which can be any other form of binary data.

Velocity: Where the data is produced at high rates

There is another important factor, which is Veracity, where the correctness of the data must be assessed. Data governance and quality processes are defined for each organization unit within an enterprise. For cross-departmental analysis, the data is not always complete and sufficient enough. The data is scattered across various sources, different data warehouses. Although automated collection exists for real-time traffic data. In certain cases, there is a limited trust being placed in someone else's data between departments. Especially in the case of the new BDA approach.

B. Database

The modern architecture requires multi-core CPUs with fast communication between processor cores, and terabytes of main memory. All data should be available in main memory, which avoids the performance penalty of disk I/O. The disk/SSD is still required for permanent persistency in the event of an outage.

A database table is organized in rows and columns and can be represented in row-order or column-order. Columnar data storage allows highly efficient compression. If a column is sorted, often there are repeated adjacent values. Compression methods can be run-length encoding, cluster coding and dictionary coding. With dictionary encoding, columns are stored as sequences of bit-coded integers. That means that a check for equality can be executed on the integers. Columnar

978-1-5090-2866-5/16/$31.00 ©2016 IEEE

storage, in many cases, eliminates the need for additional index structures. Storing data in columns is functionally similar to having a built-in index for each column. The column scanning speed of the in-memory column store and the compression mechanisms allow read operations with very high performance.

Database applications use specific interfaces to communicate with the database management system functioning as a data source running in the context of an application server.

Business continuity for the database must be part of a strategy for improving the availability of the systems. The following database vendors are most commonly used: Oracle, Teradata, IBM DB2, Netezza, MS SQL Server, MySQL, SAP Sybase.

Traditional applications use SQL to manage and query the data stored in the database.

C. Vendor Analysis

Some strengths and weaknesses have been assessed in terms of various data warehousing technologies widely available today. The following vendor's technologies are the subject of the investigation:

- SAP HANA in-memory/Sybase
- Oracle Exadata
- IBM/Netezza
- Microsoft SQL Server 2014

Architecture

Each solution is available in appliance form except MS SQL Server 2014. Oracle Exadata runs only Oracle (formerly Sun) appliance machines. The IBM Netezza is able to run only on IBM PureSystems. The MS SQL Server 2014, that includes an OLTP database engine, is available in the Windows Server operating system only. The SAP HANA is also run on an appliance, but it is an independent appliance platform. It runs Linux operating system on x86 architectures. At least 9 x86 hardware vendor partners are certified by the SAP.

SAP HANA in-memory

The SAP HANA serves large amounts of data, up-to Terabytes with outstanding performance. With respect of indicators, there are mass computing power of the entire data quantity. From the hardware' point of view, it is platform-independent, and cost-effective because it is built up of standard hardware components, appliances. The database server is flexibly scalable. The main strengths is that the access of data is at the speed of memory chips (100x compared to traditional architectures).

Oracle Exadata

It is a hybrid combination of the Oracle RAC system and Exadata Storage Servers. The storage servers have no sharing capability, hence the RAC architecture. It is important that the storage server has a smart scan capability, which can reduce the Storage Server in order to reduce the amount of data that comes from the RAC layer. Generally, Exadata is an efficient platform if the Smart Scan is used, with certain limitations due to different architectures.

Netezza

The Netezza provides a structure called zone map that provides quick access to data blocks stored in rows in the specified range of the query. The Netezza has introduced a GROOM table command, which reformats the data in order to optimize the zone map. The GROOM is equivalent to an index re-organizing job, but it re-settles data. The Netezza recently assured the possibility that more zones map to be available when the board as clustered table (cluster base table CBT) defined.

MS SQL Server 2014

SQL Server 2014 provides support for memory-based Enterprise Edition version, in which the tables can be moved manually or automatically in memory to increase performance. The solution appeared at the beginning of the year 2014, with many new features.

Online Transaction Processing (OLTP) database engine is now been integrated into the SQL Server database. Microsoft promises more efficiency from this database, which can be achieved without changing the existing application code is a 30 times increase in performance when compared to SQL Server 2012.

Faster query speed and more advanced memory-based compression in order to reach settlement resulting data warehouse construction, which includes the amendment, clustering-column indexes. Innovations automatically improve new and existing queries planning, fine-tuning and processing. Improved scalability and availability: computing scale-up to 640 logical processors and 4TB of memory and 64 virtual CPUs in the event and if one virtual machine up to 1 TB of memory in a physical environment. Improved accessibility eliminates system outage online indexing operations.

III. ANALYTICS

In order to gain a better understanding the process workflow in BDA, below there are the typical stages of the processes:

A. Collection, filtering, structuring

Data must be extracted from ERP systems data, data warehouse, spreadsheets, email messages, images, texts, etc.

Loading and ingesting require ETL tools to complete the process.

A list of unique index values from an existing dataset should be generated. The data stored in the data warehouse can be logically connected together on the basis of the natural basic keys such as vehicle type, plate no., driver, etc. or analyzing dimensions such as time of day, season, day of the week, traffic load etc.

B. Modeling, algorithms

Modeling derives actionable information from the aggregated data to get details about the database content. Business analytics models current and historical customer performance data and traits to make predictions about future outcomes and customer behaviors. These predictions can be expressed as numerical values, or scores, that correspond to the likelihood of a particular occurrence or behavior taking place in the future.

The modern solution provides in-memory processing engine that may run on a Hadoop cluster and Spark execution framework. It should be able to scale to thousands of nodes, and designed for use in large distributed clusters and for handling big data. It is built on

the Hadoop ecosystem, which provides a collection of components that support distributed processing of large data sets across a cluster of machines. Hadoop allows both structured as well as complex, unstructured data to be stored, accessed, and analyzed across the cluster.

The access of Hadoop data is made through SQL interface. A plug-in provides access to the Spark controller, and moderates query execution and data transfer. On the Hadoop side, Spark controller provides a SQL interface to underlying tables that use Spark SQL.

The use of the R open-source statistical analysis language is supported, that offers in-memory data mining capabilities for handling large volume data analysis efficiently.

Various analyses on the data are performed, including time series forecasting, outlier detection, trend analysis, classification analysis, segmentation analysis, and affinity analysis.

A range of predictive algorithms, the R open-source statistical analysis language, and in-memory data mining capabilities for handling large volume data analysis are used efficiently.

C. Visualization

The final step is to examine relevant data, perform the actual analysis and provide with the information in the format needed. Different visualization techniques are used, such as:

- Dashboards
- scatter matrix charts
- parallel coordinates
- cluster charts
- decision trees
- Web intelligence reports, mobile applications
- Top lists

D. Provisioning

Ensures that the automated transfer of the most data is performed and that the data is stored in a single repository and it is accessible via platform independent applications.

Hadoop repository may be extended to the whole organization, so that it could benefit from BDA.

The result is predictable based interventions can increase the accuracy and thereby improving customer satisfaction

IV. Case Study

In order to have an insight of the BDA, an examination has been achieved at the Budapest Transport Privately Held Corporation (BKV) specific area of traffic data.

A. About

BKV, as a Municipality of Budapest owned public transport company, serves millions of passengers a year. It operates 2 major transport divisions, the fixed-rail services as an incumbent operator (tram, metro, suburban railways) and the rubber-tired vehicle services as a major operator (bus, trolleybus) in an integrated network in the metropolitan area of about 2 million residents. In addition to these, it also operates the cogwheel railway, the funicular and the river Danube boat service, which predominantly serve tourism. The company has 2800 vehicles, some 11 thousand employees and its performance amounted to 19 billion place-km in.

Most of the systems have been producing an enormous amount of information and thanks to the recently launched Traffic Control and Passenger Information System as well as the vehicle and infrastructure modernization program, the amount of information is growing.

B. Methods

The quantity of data available at BKV has already reached the size and complexity that discovering relationships and hidden insights, using advanced predictions it needs to take advantage of the new methods. Scenario planning will require not only an analysis but also of paradigm shifts. In order to keep competitiveness, it is essential to be open at unexpected fields as well.

There is also a high demand for recent and accurate information from the systems. In the beginning of this decade there was an attempt to introduce a previous BI system that failed. Therefore, there was a limited trust being placed before a new project.

The concept was to work with business leaders to identify the value of the analysis to the business. They had to understand impacts.

C. Technology

The selected technology is based on the SAP HANA which is a real-time analytics and applications tool. It is able to analyze business operations based on a large volume and variety of detailed data. The platform is deployed as an appliance on HP servers.

D. Sound scope

- Examination of the journey time reliability
- Examination of the schedule's biases (penalty causes)
- Examination of the vehicle's utilization

E. Facts

- Analyzed data volume: ~ 230 million records, 40 GB of data
- Period: 4870 days (~ 13 years daily data)
- Unique ID's: 5.632.699
- Unique driver ID's: 11.264

F. Expectations

- High demand for recent and accurate information as previous BI attempts failed
- Faster response in reporting (especially controlling reports)
- Delivering an effective transport system
- Passenger satisfaction provided
- Improving journey experience

G. Results

- Some examples of the immediate savings realized
- Redundant internal vehicle traffic (garage traffic) detected
- Replacements due to technical faults are more predictable.
- Instead of using existing pre-defined reports various ad-hoc analyses were performed
- Users liked the flexible user interface
- Reports run much faster
- It could diminish the using of multiple spreadsheets across organization
- Better provision on time departures/ arrivals including real time information

H. Findings

- lack of input data and/or inconsistency, duplication in core basic systems
- Missing data from legacy systems
- Incompletion of existing reports

I. Sample analysis

The consequences of biased departures may consist of penalties according to the contracts for public transport.

Figure 1 shows one of the analysis results. The extent of lacking of data is important for the reliability of the analysis. No value means there is no actual data, or data is not generated. It seems that in this case only 40% of the data can be analyzed properly. It is vast importance of knowing if the biased departure is caused by internal technical problem or traffic situation. It found out that disturbance data was missing from legacy systems.

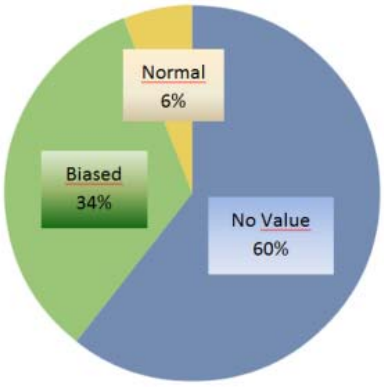

Figure 1 - *Number of biased departures*

Sample analysis also indicated that performance data in many cases are incomplete. It is crucial to make predictions about future outcomes and operators activities.

J. Aims on further research

With the advent of the recently launched Traffic Control and Passenger Information System the company is able to receive information from vehicles real-time. Not only the data derives from the GPS sensors but also there is a mobile /wireless data network installed on the vehicles. From this big source area, automated collection exists for real-time traffic data, so there is certain traffic information that are useful to the research. From this perspective it makes sense to consider Hadoop deployments and Mapreduce functions.

The aim is to build a model and examination system based on the actual and planned traffic data, so that the company will be able to examine the differences in the contexts of many different parameters, such as schedule, vehicle type, plate no., driver, time of day, season, day of the week, traffic load etc.

It is easy to foresee that it will result in more precise planning and measures that will increase accuracy, reliability and passenger satisfaction.

V. CONCLUSION

This paper briefly presented the course of developments referring to the transport management using big data analytics. Research on BDA advanced prediction techniques proves to be relevant areas of the further examinations. Because of many sensors produce vast amount of data, much more information can be known about the traffic than before. Yet due to the vast complexity of transport data, little progress has been made on the prediction so far. The data analytics helps get closer to an iterative solution and to make sensible decisions. The expected results hold out a promise to foster efficiency, cost effectiveness and improve the management of public transport.

ABBREVIATIONS AND ACRONYMS

ERP – Enterprise Resource Planning

CLOUD – Hosted services on architecture that is flexibly scalable, proprietary, public or mixed

BI – Business Intelligence

SQL - Structured Query Language

BDA – Big Data Analytics

ETL - Extract, transform, load data

HADOOP – A framework for storing data and running applications on clusters of commodity hardware

MapReduce – A framework for applications process data in-parallel nodes

R – Open-source statistical analysis language

SPARK - Fast engine for large-scale data processing

OLTP – Online Transaction Processing

GROOM – A table command used in IBM Netezza database for optimization

HANA - In-memory platform for advanced data processing and next-generation applications

SSD – Solid state drive, functions as hard drive

Repository – A directly accessible location where data is stored and maintained.

REFERENCES

[1] József Tick: Multi Server Architecture in Intelligent Traffic Control System, LINDI 2007 • International Symposium on

Logistics and Industrial Informatics • 13–15 September, 2007 • Wildau, Germany

[2] József Tick, Tamás Tiszai: Server Virtualization in Intelligent Traffic Control System, LINDI 2007 • International Symposium on Logistics and Industrial Informatics • 13–15 September, 2007 • Wildau, Germany

[3] Bőgel György: A BIG DATA ökoszisztémája, Budapest 2015, Typotex

[4] Norman Matloff: The Art of R Programming: A Tour of Statistical Software Design, No Starch Press; 1 edition (October 12, 2011)

[5] Tom White: Hadoop: The Definitive Guide, 4th Edition - Storage and Analysis at Internet Scale, O'Reilly Media Inc., 2015.

[6] Benjamin Bengfort, Jenny Kim: Data Analytics with Hadoop - An Introduction for Data Scientists, O'Reilly Media Inc., 2015.

[7] Shashank Tiwari: Professional NoSQL, Wrox, 2011.

[8] Donald Miner, Adam Shook: MapReduce Design Patterns, O'Reilly Media Inc., 2013.

[9] Chu, Wesley W. (Ed.): Data Mining and Knowledge Discovery for Big Data Methodologies, Challenge and Opportunities, Springer, 2014.

[10] Paul Zikopoulos, Chris Eaton : Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw-Hill Osborne Media 2011.