In-Database Analytics
A state-of-the-art

Annemarie Burger & Elena Ouro
Master: Big Data Management & Analytics, UPC

ABSTRACT

The explosion of data in recent years has opened a world of possibilities for companies and individuals, giving them the opportunity to derive learning from all of their data and apply it into their decision making. Traditionally the data, stored in a Database management system (DBMS) needed to be extracted and moved to a server where the analysis would take place. In-database analytics refers to the trend that seeks to integrate the analytical process into the database, making the analysis process scalable, fast and secure. In this paper we present a state-of-the-art review in the field of in-database analytics focusing on Big Data, the progress that has been made since its appearance and the current challenges researchers are facing.

INTRODUCTION

With the irruption Big Data, it became necessary to analyze much larger quantities of heterogeneous data proceeding from various different sources and stored in diverse formats. The analytical process has become more complex and sophisticated to keep up with the new data trends, the traditional approach used to perform analytics -consisting on extracting data from the DBMS where it is stored to perform the analysis in a separate system- is no longer optimal. The main issue with this approach is the extraction of data: there is no simple way to extract large quantities of homogeneous data, the memory of the system is often not large enough to contain all the data required for the analysis and it is not uncommon to have stored sensitive information that needs to be maintained properly encrypted at all times. Other factors that have pushed researchers to find alternatives to the need of shipping data to be analyzed are the need of updating all copies of the data whenever new data becomes available and the amount of time needed to extract data, as short response times are preferred when performing certain analytical tasks[1].

In-database analytics is a solution devised for this problem. The term in-database analytics refers to all analytical techniques that are applied directly to the database. In-database analytics brings into the game several advantages when compared with the traditional approach when performing predictive analytics[3]:

- **Time saving**: When using in-database analytics it is possible to take advantage of the parallel database engines to perform analytical tasks, thus, increasing the performance and reducing the response time, in contrast to performing computations using analytical tools and slow file management systems.
- **Simple processing of unstructured data**: by transforming complex statistical processes into manageable queries and moving clustering rules into the database Simple processing of unstructured database eliminates the difficulty of exporting these rules.
- **No need for temporary storage:** using in-database analytics eliminates most temporal data storage challenges since data is directly handled and stored in the database.
- **Shared environments**: in-database analytics allow business units to have access to the same raw data, data transformations and metadata. This way all collaborations have access to the most recent developments and knowledge transfer is made simple.

While "In-database analytics" usually refers to the analytic capabilities of a database- that is relational databases and column stores- the use of the term is used increasingly to refer also to the analytic capabilities embedded in data warehouses, data appliances and Hadoop clusters.[2] In this paper we will focus mainly on the database-centric model used on the analysis of big data.

TYPES

In-database analytics are performed in three main ways:

- **Translating predictive models into SQL:** This method only requires certain tasks that would be performed by the analytical tool suite to be translated into SQL queries that can natively be executed directly in the database. This is a widely used method even though it presents some language expressiveness limitations.
- **User defined functions in process space:** External libraries are loaded as user defined functions that can be called as any other function in an SQL statement. They can benefit from the database's parallelism and use its memory just like any other query.
- **User defined functions out-of-process**: Similarly to the previous, but instead of being executed in the process space of the database they are executed out-of-process which means that they have their own resources and are more safe to use but their performance won't match the one of in-process user defined functions.

USES

In-database analytics has a wide range of possible applications but it is particularly useful in systems that require the use and analysis of real-time data, and where small response times are vital, and those were sensitive data is handled and it is preferred to maintain it in the database.

A few examples of use-cases:

- **Fraud detection**: systems that need to quickly be able to execute anomaly detection algorithms over the stored data. Credit card companies for example rely on in-database analytics to detect possible fraudulent transactions, since they store years worth of usage data they can flag transactions with suspicious amounts, locations or retailers[15].
- **Energy consumption**: a study was conducted on the use of in-database analytics for the analysis of data obtained with smart meters installed in customer's houses to keep track of their energy consumption combined with the data associated with the customer, geographical location..etc[17]. By analyzing all this data they expected to identify customer consumption patterns, peaks in demand to ensure supply, possible energy theft and provide personalized feedback to customers.
- **Healthcare**: improving assisted living systems (AAL)[5], systems that "use information and communication technologies (ICT) in a person's daily living and working environment to enable them to stay active longer, remain socially connected and live independently into old age"[6]. The proposed system used in-database machine learning methods to model early night behaviours. They installed five different types of sensors in the homes of the participants in the study, whenever the sensors would insert data into the active database it would trigger a user-defined-function that would perform certain actions. They divided this actions between short term and long term, the first ones only utilized the last data inputted and often did not require any analytics, the long term actions on the other hand were those that required collecting data over a longer period of time to be analyzed to look for patterns and learn about abnormal behaviours of the subjects.

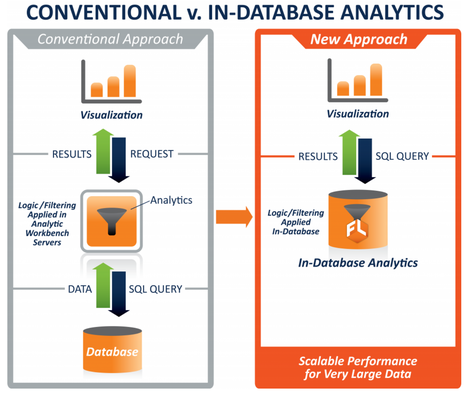
TOOLS

Querying systems, frameworks & interfaces

- **AIDA** - an abstraction for advanced in-database analytics - is a framework that keeps the computation in the RDBMS, and emulates the syntax and semantics of popular statistical Python packages in order to be able to work with both linear algebra and relational operations simultaneously. [12]
- **The Blaze ecosystem** provides an interface for multiple backends, such as SQL databases, NoSQL data stores, Spark, Hive, Impala, and raw data files [14]. This makes it easy to use, but also reduces the analytic options.
- **ibmdbpy** is an open-source Python interface developed by IBM works by connecting to distant databases and providing analytics functions that are translated in SQL queries and pushed to the RDBMS for execution, leaving the data where it is. [14]
- **SubspaceDB** is a querying system that implements subspace clustering directly within an RDBMS. It focuses on retrieving optimal answers to medoid, neighbourhood, partial similarity, and prominence queries and shows to be over 10 times faster than a conventional wrapper-based or SQL UDF approach. [16]

Database systems

- **MonetDB** is a columnar database, following the relational data model. It runs all kinds of statistical queries by just wrapping this code in a SQL user-defined function [24].
- **SAP HANA** is an in-memory, column-oriented, relational DBMS that performs in-database analytics. SAP HANA makes use of the MapReduce programming model. This supports parallelizable user-defined functions and features a generalized version of recursion to support advanced analytics. [19]
- **SAS + Teradata**: Teradata Vantage is a new "modern analytics platform.(...) SAS is an integral and essential component in a Vantage deployment, as SAS delivers analytic tools and languages that extend the new Vantage platform to create a complete analytic ecosystem." [23] Together they can deliver in-database analytics.
- **PostgreSQL + MADlib**: PostgreSQL is a great database for Big Data, since it checks the three V's: Volume, Variety and Velocity [22]. In-database analytics can be performed using PostgreSQL with MADlib. This is a free, open-source library of in-database analytic methods. It provides an evolving suite of SQL-based algorithms for machine learning, data mining and statistics that run at scale within a database engine, with no need for data import/export to other tools [21].



FUTURE WORK

We have seen the many benefits of using in-database analytics, but as with every growing technology it is not without its limitations. Researchers of the university of Wisconsin-Madison found a series of shortcomings when it came to the current use of in-database analytics while speaking to some of the leaders in the industry, Oracle and EMC Greenplum[7]. They uncovered that while in-database analytics toolkits have been available in RDBMS sine late 1990s and 2000s the main bottleneck that hindered the development of in-database capabilities in RDBMS was having to implement every new statistical technique into the system, to do so they followed an ad hoc process and every new technique would have its own set of requirements which impedes code reuse across different algorithms and makes the development progress longer. A few years later Daniel Ritter found that at present using a database-centric system showed some shortcomings in terms of language expressiveness, for example timed-aggregations were not possible, and also that when mixing SQL and PL/SQL there was some latency and the performance suffered slightly. [8]

While there exist at the present several third party libraries that offer off-the-shelf analytical algorithms, they haven't been able to keep up with some of the trends in the field. When a group of researchers identified a lacking of machine learning algorithms in the existing technologies for in-database analytics, they proposed MLog; a high-level language to integrate machine learning into data management systems.[18] However, these are only the first steps in a field that has been barely explored in the field of in-database analytics.

Much work can still be done to improve in-database analytics. On the architecture side a solution to facilitate the development of complex analytical techniques and the possibility to reuse the code between different databases is still required. In the future it is also expected that steps towards more comprehensive languages will be taken, be it by extending SQL even more or by integrating different programming languages.

REFERENCES

[1] Edouard Fouché, Alexander Eckert and Klemens Böhm. (2018). In-Database Analytics with ibmdbpy. SSDBM'18.
[2] James Taylor. (2013). In-database Analytics. Decision Management Solutions.
[3] Sue Balkan and Michael Goul. (2010). Advances in Predictive Modeling: How In-Database Analytics Will Evolve to Change the Game. Business Intelligence Journal vol.15, No.2
[4] In-database processing wikipedia entry. Retrieved on June 28, 2019 from https://en.wikipedia.org/wiki/In-database_processing
[5] Wagner O. de Morais, Jens Lundström and Nicholas Wickström. (2014). Active In-Database Processing to Support Ambient Assisted Living Systems. Sensors 2014, 14, 14765-14785.
[6] AAL programme. Retrieved on June 29, 2019 from http://www.aal-europe.eu/
[7] Xixuan Feng, Arun Kumar, Benjamin Recht, and Christopher Ré. (2012). Towards a Unified Architecture for in-RDBMS Analytics. CoRR, abs/1203.2574
[8] Daniel Ritter. (2014). What About DAtabase-centric Enterprise Application Integration?. 6th Central-European Workshop on Services and their Composition, ZEUS 2014
[12] D'silva, J., De Moor, F., & Kemme, B. (2018). AIDA. Proceedings Of The VLDB Endowment, 11(11), 1400-1413. doi: 10.14778/3236187.3236194 http://www.vldb.org/pvldb/vol11/p1400-dsilva.pdf
[14] Fouché, E., Eckert, A., & Böhm, K. (2018). In-database analytics with ibmdbpy. SSDBM. https://dbis.ipd.kit.edu/download/ibmdbpy_ssdbm18.pdf
[15] Kuchipudi Sravanthi and Tatireddy Subba Reddy. (2015). Applications of Big data in Various Fields. (JCSIT) International Journal of Computer Science and Information Technologies, Vol.6.
[16] Harikumar, S., & Kaimal, M. (2019). SubspaceDB : In-database subspace clustering for analytical query processing. Data & Knowledge Engineering, 121, 109-129. doi: 10.1016/j.datak.2019.05.003.https://www-sciencedirect-com.proxy.uba.uva.nl:2443/science/article/pii/S0169023X17305633
[17] Xiufeng Lu and Per Sieverts Nielsen. (2015). Streamlining Smart Meter Data Analytics. 10th Conference on Sustainable Development of Energy, Water and Environment Systems.
[18] Xiupeng Li, Bin Cui, Yiru Chen, Wentao Wu and Ce Zhang. (2017). MLog: Towards Declarative In-Database Machine Learning. VLDB Endowment 2150-8097/17/08.
[19] Carsten Binnig, Norman May, Tobias Mindnich (2013). SQLScript: Efficiently Analyzing Big Enterprise Data in SAP HANA. In: Markl, V., Saake, G., Sattler, K.-U., Hackenbroich, G., Mitschang, B., Härder, T. & Köppen, V. (Hrsg.), Datenbanksysteme für Business, Technologie und Web (BTW) 2035. Bonn: Gesellschaft für Informatik e.V., (S. 363-382).
[21] Hellerstein, J., Li, K., Kumar, A., Ré, C., Schoppmann, F., & Wang, D. et al. (2012). The MADlib analytics library. Proceedings Of The VLDB Endowment, 5(12), 1700-1711. doi: 10.14778/2367502.2367510
[22] Blitz, S. (2018). Is PostgreSQL the Right Reporting Database for You? | Sisense. Retrieved from https://www.sisense.com/blog/decide-postgresql-reporting-right/
[23] Heather Burnette, Greg Otto, Salman Maher (2019). SAS and Teradata: See the Advances. SAS, Paper 3652-2019. https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2019/3652-2019.pdf
[24] Boost your Data Analytics. MonetDB solutions. https://monetdbsolutions.com/solutions/analytics
Picture: http://www.fuzzylogix.com/products-old/db-lytix/