Big Data Science and EXASOL as Big Data Analytics tool

Ajit Singh, Sultan Ahmad, Mohammad Imdadul Haque

Abstract**: *Big data and Data science are the two top trends of recent years. Both can be combined together as big data science. This leads to the demand for new system architectures which facilitates the development of processes which can handle huge data volumes without deterring the agility, flexibility and the interactive feel which suits the exploratory approach of a data scientist. Businesses today have found ways of using data as the principal factor for value generation. These data-driven businesses apply a variety of data tools as data analysis is one of the chief elements in this process. In order to raise data science to the new computational level that is required to meet the challenges of big data and interactive advanced analytics, EXASOL has introduced a new technological approach. This tool enables us more effective and easy data analysis.***

Index Terms**: *Big Data Science, EXASOL, Big Data, Data Science.***

I. INTRODUCTION

Big data which involves enormous data sets, has redefined the role of data in businesses. The capability to provide for sophisticated analytics, integrating diverse existing data sources is essential for the success of today's businesses. Today, all kinds of data are collected all the time by everyone and with the current adoption of the Internet of Things in mind by everything. The importance of data and its relation with business application can be shown in Fig.1.

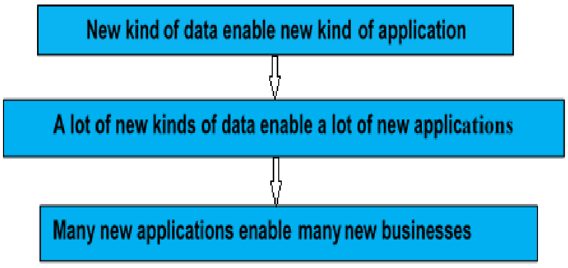
New kind of data enable new kind of application
A lot of new kinds of data enable a lot of new applications
Many new applications enable many new businesses

Fig. 1: Relation of data, application and business

Businesses today have found ways of using data as a most important factor for value generation. These data-driven businesses apply diverse data tools to perform data analysis. This data analysis is one of the major components of this process. Consequently, over the last few years data science has developed into a thriving profession receiving enormous attention from both industry and education. Briefly, data scientists explore, integrate, model, evaluate, and automate, thereby helping to create processes, which generate value from data.

Beyond the actual tools and details of the data, it is the data science way of working which is crucial. R and Python allow us to interact closely with data, to get our hands dirty, play with data in an ad-hoc fashion; and this way of working needs to be preserved when moving to larger scales.

Data scientists are in a disadvantageous position when huge amount of data is involved as large-scale data analysis tools are too rigid to bear the data analyst's approach of working. Michael Stonebreaker, the winner of the Turing Award 2014, has expressed "The change will come when business analysts who work with SQL on large amounts of data give way to data scientists, which will involve more sophisticated analysis, predictive modeling, regressions and Bayesian classification. That at scale does not work well on anyone's engine right now".

The consequence is obvious: As data scientists cannot work with very large data sets interactively, they fall back to two compensation strategies. Either they are working batch-oriented and lose major components of their unique style of working that makes them so powerful or they are working only on small subsets of the real data. In general, the latter strategy is fine for some data exploration activities, but it means that putting new insights into production remains sluggish. In response to the two developments of recent years - big data and data science, EXASOL provides for a new systems architecture for big data science. It allows the consumer to carry out complex analytical tasks on huge amount of data in an interactive manner, right in the database, by means of several programming languages.

II. LITRATURE REVIEW

Big Data services rely on the proper analysis and management of huge business data. Data scientists 'Chen and et al', while reviewing the issues related to Big data with a management perspective observed an extraordinary growth in the volume and nature of data of diverse entities like Web companies, enterprisers, physical and science researchers. Traditional data processing processes are insufficient due to the huge volume of data available [1]. The study identified Data integration, Data reduction, Data querying and indexing, Data analysis and mining for in-depth analysis.

Another scholar 'Kalra and et al' identified the

Revised Manuscript Received on July 12, 2019.

Ajit Singh, Patna Women's College, Patna University, Patna India.
Sultan Ahmad, Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, AlKharj, Saudi Arabia.
Mohammad Imdadul Haque, College of Business Administration, Prince Sattam Bin Abdulaziz University, AlKharj, Saudi Arabia..

challenges of big data as "timeliness and heterogeneity, scalability, performance, continuous availability, workload diversity, data security, identifying right data, identifying right talent, identifying right platform, identify right architecture, collaborating across functions and businesses" [2]. Also, in terms of data visualization the study identified the following challenges: "meeting the need for speed, understanding the data, addressing data quality, displaying meaningful results and dealing with outliers".

Jaseena and David state that data is an essential element of "economy, industry, organization, business function and individual" [3]. Big Data has distinctive computational and statistical challenges of "scalability and storage bottleneck, noise accumulation, spurious correlation and measurement errors" for which "new computational and statistical paradigm" are required. While introducing some data mining tools the study opines that big data is central for automatic discovery of intelligence present in patterns but are hidden. This allows business in decision making, forecasting and identifying new changes and opportunities.

In another study, Fan and et al stated that Big Data analytics discovers "subtle population patterns and heterogeneities" but is prone to "computational and statistical challenges, including scalability and storage bottleneck, noise accumulation, spurious correlation, incidental endogeneity and measurement errors" [4]. The study cautioned against inability to validate the exogenous assumptions owing to incidental endogeneity which may result is faulty statistical inferences.

Mandale, A. and Gadage, S. opine that Big Data analytics is applicable to all dimensions namely "mobile services, retail, manufacturing, financial services, life sciences, and physical sciences" and faces issues of "heterogeneity, scale, timeliness, complexity, and privacy" [5].

Scholar 'Jin and et al' in 2015 stated that Big data has quickly established into a burning topic that appeal to wide-ranging consideration from academics, industry professionals, and governments [6]. The study categorized the structures of big data as 5Vs, as high Volume, Velocity, Variety, value and low veracity. The study identifies the challenges to big data as "data complexity, computational complexity, and system complexity". Big data is significant for national development, industrial upgrades, scientific research, interdisciplinary research, help in perceiving the present and predicting the future. The study identifies the challenges to big data as Data complexity, Computational complexity and System complexity and finally calls for advances in processes to handle the data better

Kubina and et al opines that to performance in a competitive environment depends on the ability tom process relevant information at the appropriate time [7]. The study emphasized that vast quantities of information is present in various forms. The challenge is to have new processes and technologies to process the information to the competitive advantage of the business. As an example, simple data from each transaction can be used for tracking and predicting demand as it can give insights to consumers buying behavior leading to tailor made service and products. Its advantage over traditional methods of analysis in terms of processing unstructured or semi structured data without going through the issues of sampling framework at a lower storage cost.

Pai states Challenges, Tools and Techniques needed for Big Data. Big Data requires new and advanced methods for "managing, analyzing, visualizing, and exploiting informative knowledge from large, diverse, distributed and heterogeneous data sets" [8].

Mohan in 2016 states that Big Data is redefining management principles and is now the "new fact of business life" [9]. Future benefits to companies would depend on the usage of big data through suitable and advances analytical processes. The study also identifies the benefits of big data in terms of saving costs, competitive advantage and fresh business prospects. The study discusses the benefits of big analytic in the field of financial institutions, governments, telecommunication and marketers. Finally the study calls for fresh innovations in data analytics processes.

Schroeder in 2016 suggested that opportunities exist to improve upon the collection, storing and analyzing of data. The progress so far in big data analytics, is neither a 'big bang' type nor simultaneously in all sectors [10]. Contrarily, it is a gradual evolution. There is an ample scope for additional businesses to avail the benefits. The study acknowledges internal challenges for big data forms in terms of procedural issues related to collection, storing and processing of data.

Jelonek in 2017 opined that big data analytics provide for the management of businesses through uncovering unknown patterns, unknown correlations and trends in market [11]. These improve operational efficiency, competitive advantage and ultimately profitability of businesses. The study identified the advantages of big data analytics over traditional analysis in terms of unstructured formats of data, enormous volume and constant flow data with machine learning methods as compared to row/column format of static data using hypothesis based analysis.

Alsghaier and et al. in 2017 while advocating the usage of Hadoop, an open source platform, for processing data, termed the developments in big data as revolutionary transformation to attain competitive advantage in terms of business performance [12]. The researchers recognized that big data is still in its developmental; phase as is prone to programming issues and development of platforms like Hadoop can facilitate big data analytics.

Tukkoj and Seetharam are of the opinion that Big Data is "catalyzing new business models and reshaping industries" and the need is to develop advanced platform to analyzing high volume data for variety and velocity[13].

Recently in 2018, Naganathan is of the opinion that Big Data "has the potential to revolutionize the art of management to take appropriate decision on time" [14]. Big Data analytics "reveal patterns, trends, and association from unstructured data into structured ones to find a solution for a business". The study further discusses methods applied by various organizations.

Rahaman and et al. in 2018 acknowledged that traditional methods cannot be used to current data sets because of its huge volume, velocity and variety [15]. The study expects that the volume of data extracted through different applications would double every two years. That's a major reason why new and advanced techniques are required. Finally, the study recommends further research on new tools of data analysis

III. NEW TECHNOLOGY APPROACH FOR BIG DATA

In order to raise data science to the new computational level that is required to meet the challenges of big data and interactive advanced analytics, EXASOL has introduced a new technological approach.

A. Advancement in Memory Technology

Advanced in-Memory technology is required for achieving the required performance goals and can be shown in Fig.2. This supports in-memory operation on datasets that are larger than physical RAM without major decreases of performance.

- Massive Parallel Processing is required for providing the scalability that is necessary for any big data application. In particular this means that all key-operations are performed in parallel on every available physical machine, yielding maximum performance.
- In-database analytical programming is necessary for providing the required flexibility to allow data scientists to bring the algorithms to the data.

B. EXASOL as new strategy for Big Data

EXASOL is a good starting point for pursuing a big data science strategy. At its core, EXASOL is an analytic relational database management system with all the features anyone would expect such as transactions, multi-user support and compliance with important industry standards. It delivers the best in-memory technology and the best massive parallel processing on the market. Those features have made EXASOL the unrivalled champion of the TPC benchmark H for ad-hoc decision support for many years now. In order to move more and more computation to the data, EXASOL is constantly adding new features for in database analytical programming to the core system. For instance, it is possible to use any programming language for creating user defined analytics which will be executed fully in parallel and distributed in the EXASOL cluster for maximum analytical performance.

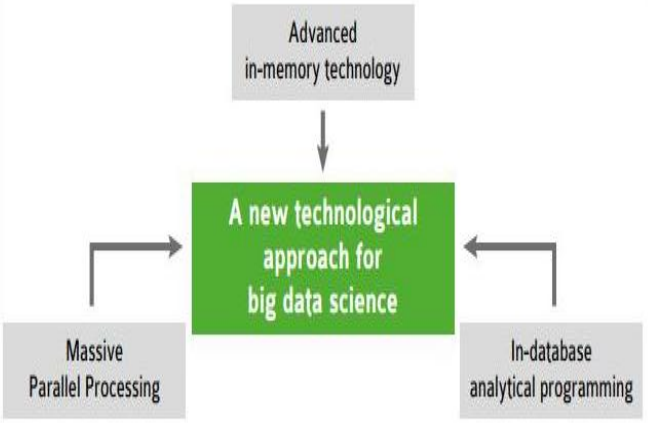

Fig. 2: New Technologies approaches for Data Science

IV. HOW TO START WITH BIGDATA ON EXASOL

For data scientists who are using R or Python, the easiest way to get started with EXASOL in-database analytical programming is to start from where they stand at the moment. EXASOL provides packages for R and Python that allow you to run code in EXASOL directly from the R and Python command line. Let’s have a look: First the package (and the RODBC package) needs to be loaded:

```
> require(exasol)
```

First, create a connection to the database:

```
> C <- odbcConnect(„ your Data Source Name “)
```

Then, if we want a function from our R session to be available in EXASOL, we simply deploy it in EXASOL with the exa.createScript command. For instance, to deploy our own version of the MEAN function using R, we issue a command like this:

```
testscript <- exa.createScript( C, „ test.mymean
 “, function(data) {
data$next_row(NA); # read all values from this
group data$emit(mean(data$val)) },

inArgs = c(„ groupid INT “, „ val DOUBLE “),
outArgs = c(„ groupid INT “, „ mean DOUBLE “))
```

This call to *exa.createScript* does two things:

(1) It deploys the supplied anonymous function (of course this function could also be named) in EXASOL using the name test. mymean.

2) It returns a new R-function (which gets assigned the name testscript) which transparently causes the execution of test. mymean in EXASOL on the supplied data- and data grouping specification. In order to execute this function in EXASOL, we simply type:

```
testscript(„ groupid “, „ val “, table = “test.
twogroups “, groupBy = „ groupid “)
```

val and groupid are the names of the columns which are to be fed into test.mymean. The table parameter defines the data which is fed into test.mymean.table and can be defined by arbitrary complex SQL statements. Finally, as test.mymean defines an aggregate function, we use the groupBy parameter to define the SQL groups which we would like to aggregate.

As previously discussed, the custom aggregate function test.mymean from the previous section has been deployed right in EXASOL where it is available for inclusion in standard SQL queries. The following SQL queries corresponds to the previous call to testscript:.

SELECT test.mymean(groupid,val) FROM test.twogroups GROUP BY groupid;

mymean can also be created right from the SQL interface:

CREATE R SET SCRIPT test.mymean(groupid INT, val DOUBLE)
EMITS(groupid INT, mean DOUBLE) as run <- function(data) {
    data$next_row(NA); # read all values from this group

```
data$emit(mean(data$val))
};
```

In addition to functions that integrate right into SQL, EXASOL offers a scripting facility that issues arbitrary SQL queries and then incorporates their results into a complex workflow.

V. TECHNICAL DETAILS

In this section, we discuss the technical details that are at work behind the scenes. But first we need to introduce some terminology. In EXASOL users can create user defined function scripts (UDFS) of various kinds in different programming languages. These scripts can be used freely inside SQL queries (inside SELECT statements) where depending on their type and context of use, they may serve the purpose of user defined functions (they return values) but also as generators (they return many rows) and shown in Fig.3.

UDFS are executed in parallel and distributed across all machines in the EXASOL cluster.

On the other hand, EXASOL Scripting is a single- threaded programming language that allows users to issue queries to the database. Scripting has to be invoked by EXECUTE statements. Scripting typically is used to orchestrate a number of SQL queries.

.

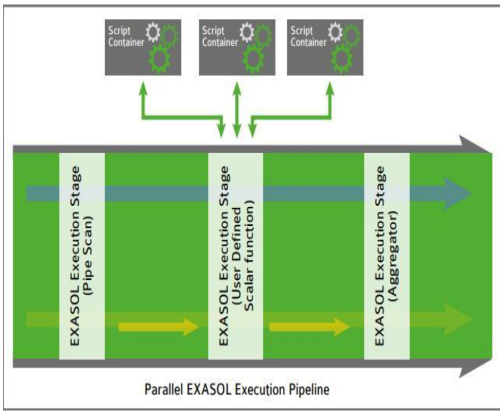
Script Container
Script Container
Script Container
EXASOL Execution Stage (Pipe Scan)
EXASOL Execution Stage (User Defined Scalar function)
EXASOL Execution Stage (Aggregator)
Parallel EXASOL Execution Pipeline

Fig. 3: Exasol Execution process

In EXASOL, the execution of user defined function scripts is tightly integrated with standard SQL execution. In other words, in order to run a user defined function, it has to be part of a SQL query. When the SQL query executes, the user's code is run. User defined function scripts take advantage of EXASOL's advanced in-memory technology and its parallel and distributed programming capabilities.

Therefore, it is not too uncommon to have a situation where hundreds of instances of a script are running in parallel on an EXASOL cluster. They are benefiting from the system's massive computational power in order to get maximum performance.

While the execution is tightly integrated with EXASOL SQL processing, the actual program code run is encapsulated using container technology. Such containers are managed by EXASOL:

- EXASOL decides how many script containers are needed on each machine and creates their running instances.
- EXASOL also decides which resource limits are applied for each script container, in such a way that a misbehaving script container is not able to block all available resources, especially since EXASOL contains full multi-user support.

The isolation of the EXASOL database from the script containers allows for secure script programming. After a script container is started, it communicates with EXASOL via an open network protocol. Connections are established via ØMQ5 sockets and messages are encoded using Google's protobuf6 library.

By nesting multiple SQL queries containing user defined functions (for instance via subselects), a large number of user defined functions can be executed at the same time, where data flow between the function script containers is ensured via the EXASOL's SQL processing pipeline. This pipelining SQL execution engine is able to engage many script containers and other SQL processing components in parallel without creating costly materializations of the processed data.

As the internals of script containers are completely isolated from the core database and the communication is performed via network messages, there are no constraints on how script containers operate inside. Indeed they can be created using any language and technology, as long as they are able to talk to EXASOL using the aforementioned open protocol.

VI. CONCLUSION

Big data science will revolutionize the way businesses generate value from data. It provides the ability to create, deploy, and interact with production quality data science models right where the data is stored. In addition, by wrapping big data science in a standard SQL interface, EXASOL provides a smooth transition from traditional BI to big data science, both for analysts and for their SQL toolsets.

In this paper we have discussed how big data science architectures result from the convergence of the following technologies: advanced in-memory, massive parallel processing, and in- database programming. This is the very reason why EXASOL is the perfect solution if anyone wants to build and create an agile and scalable big data science system.

REFERENCES

1. Chen, J., Chen, Y., , Xiaoyong D.U., Cuiping L.I., Jiaheng L.U. ,Suyun Z., Xuan Z., "Big data challenge: a data management perspective", *Front. Comput. Sci.*, vol. 7(2), 2013, pp. 157–164.
2. Kalra, B., Yadav, S., and Chauhan, D.K. . "A Review of Issues and Challenges with Big Data*", International Journal of Computer Science and Information Technology Research*, vol. 2(4), 2014, pp.97-101.
3. Jaseena K.U and David, J.M., "Issues, Challenges, and Solutions: Big Data Mining", *Computer Science & Information Technology*, 2014, pp. 131–140.
4. Fan, J.,Han, F., Liu, H., "Challenges of Big Data analysis . National Science Review", vol. 1, 2014 , pp. 293–314.
5. Mandale, A. and Gadage, S., "Big Data Analytics: Challenges, Tools". *International Journal of Innovative Research in Computer Science & Technology*, vol. 3(3), 2015, pp. 10-14.
6. Jin, X., Wah, B.W., Cheng, X., and Wang, Y., "Significance and Challenges of Big Data Research", *Big Data Research*, vol. 2, 2015, pp. 59–64.
7. Kubina, M., Varmus, M., and Kubinova, I., "Use of big data for competitive advantage of company" *Procedia Economics and Finance*, vol. 26, 2015, pp. 561-565.
8. Pai, V. "Big Data New Challenges, Tools and Techniques", *International Journal of Engineering Research and Modern Education,* vol. 1(1), 2016, pp. 743-750.

9. Mohan, A. Big Data Analytics: Recent Achievements and New Challenges. *International Journal of Computer Applications Technology and Research*, vol. 5(7), 2016, pp. 460-464.
10. Schroeder, R. "Big data business models: Challenges and opportunities". *Cogent Social Sciences*, vol. 2, 2016, pp. 1-15.
11. Jelonek, D. "Big data analytics in the Management of Business". *MATEC Web of Conferences,* vol. 125, 2017.
12. Alsghaier, H., Akour, M., Shehabat, I., and Aldiabat. S., "The importance of big data analytics in business: A case study", *American Journal of Software Engineering and Applications,* vol. 6(4), 2017, pp. 111-115.
13. Tukkoji, C and Seetharam, "A Comprehensive Survey on Big-Data Issues, Challenges and Management Approaches on Cloud Environment", *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 6(2), 2017.
14. Naganathan, V, "Comparative Analysis of Big Data, Big Data Analytics: Challenges and Trends", *International Research Journal of Engineering and Technology,* vol. 5(5), 2018, pp. 1948-1964.
15. Rahaman.A., Rajesh. S, Rani, G., "Challenging tools on Research Issues in Big Data Analytics" *International Journal of Engineering Development and Research*, vol. 6(1), 2018, pp. 637-644.
16. https://www.exasol.com/en/community/resources/resource/a-peek-under-the-hood/ Whitepaper: EXASOL :A Peek Under The Hood
17. https://www.exasol.com/en/community/resources/resource/a-drill-down-into-exasol/ Whitepaper : A Drill-Down into EXASOL

AUTHORS PROFILE

Ajit Singh is working as Asst. Professor (Ad-hoc) in Department of Computer Application, Patna Women's College, Patna University, Patna, Bihar. He is also a PhD candidate at Patliputra University, Bihar, India working on Social Media Predictive Data Analytics at the A. N. College Research Centre, Patna, India. He holds M.Phil. Degree in Computer Science, and is a Microsoft's MCSE / MCDBA / MCSD. He has 20 Years of strong teaching experience for Under Graduate and Post Graduate courses of Computer Science across several colleges of Patna University and NIT Patna, Bihar, India.

Sultan Ahmad is currently working as Senior Lecturer in Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al- Khari, Saudi Arabia. He has a unique blend of education and experience. He has received his Master of Computer Science and Applications from the prestigious Aligarh Muslim University, India with distinction marks. He has graduated in Computer Science and Application in 2002 from Patna University, India. His research and teaching interests include Cloud Computing, Big Data, Fog/Edge Computing and Internet of Things. He has presented and published his research papers in many national and International Conferences and in many peer-reviewed reputed journals.

Mohammad Imdadul Haque is an Associate Professor and Head of Management Department, College of Business Administration at Prince Sattam Bin Abdulaziz University, Kingdom of Saudi Arabia. He is a Ph.D. in Economics from Aligarh Muslim University, India. He has around twenty-five research papers, over six conference presentations and another six funded university projects to his credit. He has a rich experience of using software like SPSS, Stata, and Eviews for both multivariate analysis and econometric analysis.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication