Soaring Performance Analytics Intelligence in Big Data Approaches

S.Harini Krishna
Assistant professor, Dept. of CSE
Geethanjali college of Engineering and Technology

K.Gnana Mayuri
Assistant professor, Dept. of CSE
Geethanjali college of Engineering and Technology

Abstract: ***This thesis explains Big Data Phenomenon, which is characterised by rapid growth of volume, variety and velocity of data - information assets, and thrives the paradigm shift in analytical data processing. Thesis aims to provide summary and overview with complete and consistent image about the area of High Performance Analytics (HPA), including problems and challenges on the pioneering state-of-art of advanced analytics. Overview of HPA introduces classification, characteristics and advantages of specific HPA method utilising the various combination of system resources. In the practical part of the thesis the experimental assignment focuses on analytical processing of large dataset using analytical platform from SAS Institute. The experiment demonstrates the convenience and benefits of In-Memory Analytics (specific HPA method) by evaluating the performance of different analytical scenarios and operations.***

Keywords: Big Data, advanced analytics, in-memory analytics, in-database analytics

1. INTRODUCTION

The world has turned into information society that highly relies on data. Since information systems generate enormous amounts of records every day, every second, it seems the world is reaching the level of data overload. It is obvious now, that in order to process such volumes of data an enormous capacity is required in terms of storage and computing resources. Whereas the growth of capacity is limited by evolution of hardware and technologies, the growth of the data volume is in fact unlimited.

Getting more specific, nowadays many organisations has adopted and broadly use information systems running on technological platforms, many their agendas has become addicted to data. In mature organisations data directly affect the logic of business processes, information has become a core of their business or business end. Hence business demands the data, furthermore availability of specific data in specific time. More and more complex and risky decision making process relies on correctness and transparency of data. 1.1 Motivation Interesting driver related to this topic mentions that the growth of data is unlimited. What is the society going to do about the data overload? How to handle and moreover to process all the data? Seems like we are having the Big Data issue.

Another driver for this topic is retrieving the information (not to gather all data for further analysis). Among all the data, how to retrieve the relevant information and within a required time? Which analytics should be applied on data? What is the balance between cost of retrieval and value of that information? What are the costs of capacity to retrieve desired information? It seems like it is all about the profit, trade-off between value of information and the cost to get it. Additionally to both drivers the challenge is to visualise the information in such a way that its value is comprehensive and understandable. The main issue is the information overload.

Analytics in the traditional mode, in terms of the Big Data, are acquiring data that may or may not be needed for analysis. This all requires an innovative point of view, a different approach, architecture or infrastructure, if any. High performance analytics is one of them.

Adopting new technologies requires to process, discover and analyse these massive data sets that cannot be dealt with using traditional databases and architectures due to the lack of capacity resources in terms of computation and storage. High performance analytics represents one of the innovative approaches that can be applied on the increasing volumes, velocity and variety of data. 1.2 Goals Big Data Phenomenon, which is characterised by rapid growth of volume, variety and velocity of data - information assets, thrives the paradigm shift in analytical data processing. High Performance Analytics (HPA) can be considered as one of the approaches. The aim of the thesis is a research (overview, classification, discussions on problems and challenges) on the pioneering state-of-art of advanced analytics utilising various methods (HPA methods) that could escalate and optimise the computation performance of analytics.

Considering the fact that the selected area of research is currently being refined and formalised and simultaneously is emerging rapidly in proprietary definitions and solutions from multiple vendors, the goal of the thesis is to classify and provide summary and overview with complete and consistent image about the area of High Performance Analytics. Moreover, utilisation of these methods shall be demonstrated in practical assignment involving a processing of huge dataset.

2. BIG DATA PHENOMENON

As it has been mentioned before, the main raw material in this topic are data, Big Data. In this section, the Big Data Phenomenon is approached from its starting points and causalities. As the growth and volume of data appeared as a remarkable problem for capturing, handling and processing, it has been reactively described by many authors.

Later in this section, a definition of Big Data is provided (as it was initially described by Gartners). Additionally, Big Data can be characterised by four dimensions (4V). Dimensions are used to materialise notion of the Big Data.

Once the phenomenon is defined, discussion about its influences and impacts may start. How does Big Data influence the information infrastructure and technologies? What are the impacts on data storing and data processing? How are the processes and architecture of information and analytics systems affected? Where are the trade-offs and dependencies between operational problems and risks on one side, and innovative possibilities and opportunities on another?

Finally, the missing part to the phenomenon is the way how to handle Big Data. That is evolving questions: Are there already existing solutions to solve the Big Data Phenomenon? What are the architectures of these solutions? Which technologies are used, or which has been designed for this purpose?

2.1 Starting points: How did everything start? Early in 80's and 90's, the first information systems (IS) started exploiting in enterprises and organizations across various industries. Information

systems slowly generated more and more data. Enough data that the sense to examine, search and analyse them for information become genuine. Information that could unveiled trends, dependencies, causalities and hidden patterns.
Generated data remained untouched until the evolution of advanced information systems reached the maturity level to be able to effectively process data with analytic methods. Availability of memory (especially with direct access, random access memories) and computation power had been rapidly evolving.
Considering both starting points, enough data and enough data processing capacity, what are causalities of this phenomenon?

2.2 Causalities Let's first look at the causalities in data processing perspective. Many vendors came up with solutions, mostly monolithic, that were able to setup infrastructure for collecting and managing data. Main idea was to pump everything (generated data) in one central storage. So called Data Warehousing (composed of storages: Data Warehouse, Data Marts) became a core paradigm as a main method and technology for extracting data. Many big sized enterprises and companies adopted it hoping that it would bring them insights and overviews of data. Data Warehouse should have contained all information needed, the one version of the truth, and make it available for business analysis and decision making.
Understating the goal However, what if business data changed the form, format, syntax, semantics, what would happen then with Data Warehouse solution to which all data had been pumped up? Are there any needs for its extension, modification, recreation from scratch? Projects with building Data Warehouse have become sprints on long-term runs.
But what if the data continue growing and accumulating, how would the analytical methods applied on data with unremitting growth perform? Furthermore, business requirements on results of analysis have been shifting along with the alterations in business.
Initially, consideration of goal was in extracting, transforming and loading data into central point and consequently utilizing analytical processing to unveil information for business users. Later, the goal was to process all data. IT managers were contented. What about business managers? Everything in this process seemed to be cool – infrastructure, technologies, processes, etc. Except that it was not meant to build an information fortress. It the end, business users seek for concrete information that can be converted into added value for business, competitive advantage and profitability. Having accurate and valuable information has been identified as the crucial target.
Information boom Phenomenon is powered by multiple aspects related to data and/or information. Let us discuss further causalities, such as sources, structures, growth and relevance of data and/or information.
IT systems have become broadly and enormously incorporated in our daily life. All systems, devices, software that generate any data can be considered as Data Sources. Internet as unlimited data source (text, multimedia content: pictures, video, documents), mobile technologies (phone calls registry, messaging, location sampling), systems in any industry, transportation, medicine, services, government, trading (trades, prices moves, bids and offers, clearing data) generate data. Furthermore, some data, let us call them primary, generate another secondary data. Information is shared nowadays faster than human can even realise. Supply of data has exceeded their demand.
Apart from various sources, data are generated and offered in variety of different Data Structures. Just imagine melange of data in different forms and formats; with different syntax and semantics; having different life cycles and quality. Thinking about unique and monolithic solution for supporting all kinds of data is almost impossible.
Speaking of Data Growth, is there any limit how far the data might breed? In 2012, the 2,5 exabytes of data were emitted every day in 2012 according to estimations by IBM [12], roughly 1 zettabyte of data. This rate almost doubled in last 2 years, mostly affected by massive use of social networking (400 million tweets a day on average, 500 million daily active users on Facebook), spread mobile technologies (4 billion of mobile phone in use). And it is not only a volume, the complexity of data grow as well.
Among all those data, the importance of extracting the information has become substantial, e.g. for organisations to take a competitive advantage, in medicine to explore and predict diseases, etc. Data Relevance has become a challenge for anyone who wants to create the most beneficial and profitable information that could be absorbed and effectively utilized by users. These causalities stress the necessity of definition of Big Data.

3. APPROACHES TO HANDLE BIG DATA

Big Data can be seen as problem, on the other hand as opportunity (see previous section). This section contains approaches of handling the relative big data from perspective of architecture, processes, infrastructure, and technologies.

3.1 Approaches: The traditional BI architecture can be considered as a starting point for architectures with its process (including staging area, data-warehouse, data marts, ETL, etc.). Looking for limitations of this architecture, one may find out that it is unlikely to store all data in central (enterprise) data-warehouse and not all data are necessary to be stored. [10]
There have been new architectural approaches evolved: Hybrid Storage Architecture (combination of storages for various data types and formats, temporary data storages, data stream processing), Upstream Intelligence (analytical and statistical functions are applied early in the process during acquisition of data) that includes also specific Stream and Event Processing (based rule-based systems, pattern identification).
As results of this evolution, Post-modern BI Architecture represents a complex solution that has been inherited, mostly from traditional Business Intelligence and adds the concept of hybrid storage architecture, upstream intelligence, and stream an event processing. Postmodern BI Architecture consists of distributed data-warehouse, consolidated meta-data layer, coordinated management of data streams (ETL) and collaboration knowledge management.
Hybrid Storage Architecture Hybrid Storage Architecture represents an idea of combining a traditional data-warehouse (designed for structured data) with various data storage for different types and structures of data and analysis applications (Hadoop, NoSql). The concept also considers temporary data storages that buffers data, additionally data can be discarded after the certain time-out. Critical information (content and time wise) can be stored on platforms with advanced technologies, directly supporting analytical processing with scalable performance. The less relevant information or not time critical information can use simpler and cheaper technologies. To sum it up, different destination for different data is determined according data priorities and further processing methods.
Upstream Intelligence Idea of Upstream Intelligence is simple: apply analytical processing and methods (typically included as a last bit of process - downstream) in the initial phase of extracting data from data source - upstream. The aim is to control data that go into data management systems by filtering out irrelevant information, hence data can be evaluated immediately (by using analytical and statistical functions) accordingly to the importance and relevance prior to being stored.
Technically, analytics can be plug into the ETL processes (user defined analytical processes can be deployable in upstream). This allows to immediately analyse quickly aging data (from perspective of business). Therefore Upstream Intelligence can be wired with the processing of data streams and events, which is implemented by Complex Event Processing (CEP) technology. CEP supports

continuous intelligence and they are used as intelligent sensors that can be attached to streams with large volume of data and monitor combination of events in (near) real time.

3.2 Post-modern: BI Architecture Due to diversity of requirements from business orthogonal BI architectures evolved: a TopDown and a Bottom-Up architecture. The Top-Down architecture stressed out a reportdriven or a data-driven approach where a data warehouse model is created first based on the business/reporting requirements. Process of this approach starts with an ETL routine to move data from source system to the data warehouse (DW), and then continues with creating reports and dashboards to query data in DW. This approach mostly satisfies casual users with periodical reporting and monitoring. [7]

Apart from that, organisations pursuits power users to work on ad-hoc analysis or tasks in research and development department. With previous approach power users are left aside to use ad-hoc spreadsheets, separate/local database instances, SQL and data-mining workbenches. With Top-Down approach power users find BI tools inflexible and a datawarehousing structure too limited for their concerns. Opportunity for Bottom-Up architecture approach has appeared.

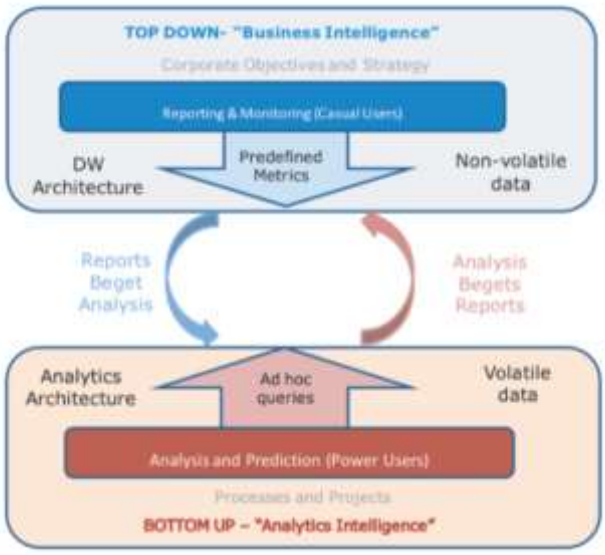

Figure 1: Top Down vs. Bottom Up architecture approach. [6]

The Bottom-Up approach suits better for business analysts and data scientists who require the ad-hoc exploration of any data source, both inside and outside corporate boundaries, working closely with business managers to optimise existing processes. [6]

Post-modern BI architecture is a result of expansion of data warehousing architectures, data governance programs and adding advanced analytics in order to balance the dynamic between top-down and bottom-up requirements. This architectural concept is also known as Hybrid architecture (depicted in the Figure 6). Big Data and HPA do not change data warehousing or BI architectures. They simply supplement them with new technologies and access methods better tailored to meet the information requirements.

Hybrid architecture can optionally contain following complementary technologies (described more detailed in Section 2.3.2) like:

- Hadoop clusters – to support storage for semi-structured data, used in staging area or analytical sandboxes
- Streaming and Complex Event Processing Engines – to support continuous intelligence, used as intelligent sensors that can be attached to streams with large volume of data and monitor combination of events
- Analytical Sandbox – to boost analysis processing, ad-hoc queries, to satisfy shortterm analysis needs, used as an access points for other BI systems
- Non-relational database system – to store unstructured or raw data, used in analytical sandbox, or staging area
- Data hub – to feed other systems and applications rather than to host reporting or analysis applications directly, data-warehouse used as a hub

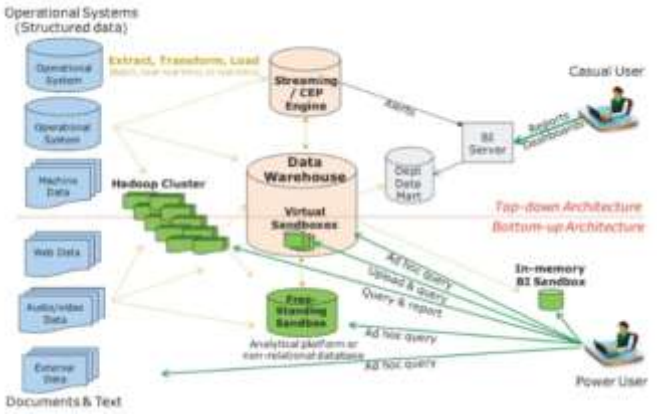

Figure 2: Post-modern BI architecture. [6]

4. HIGH PERFORMANCE ANALYTICS

This section explores the world of analytics, high-performance computing and luckily combinations of both. This section describes the drivers, definitions, classifications, technologies of high-performance analytics.

Drivers and Boundarie:s High-performance analytics (HPA) emerges with increasing demands for supporting advanced analytics and shifting paradigms about data management. Advanced analytics often demand larger, detailed data volumes than the smaller pools of aggregated data frequently used for BI and online analytical processing (OLAP). Data availability becomes crucial particularly when analytic models need to be deployed and worked against real-time or at least very timely data and traditional approaches to data integration, data management, and (ETL) processes may not be optimal for advanced analytics.

Decreasing cost of memory - RAM, SSD, hard drives (price per gigabyte), increased addressable space for memory (64bit operating system), increased computational power of hardware, increasing volume of data (Big Data), demand for real-time data processing, demand for complex analytics. All of them can be considered as drivers towards the concept of high-performance analytics.

The aim of HPA is to improve and facilitate an effective allocation of computation and the capacity resources (processing speed, computation complexity, data storage, network traffic).

Analytics There are four types of analytics. Descriptive analytics explore current trends, customer behaviour, relations, trying to focus on output parameters and to answer question "What is happening?" Diagnostic analytics explore causes and reasons within the context, trying to focus on input parameters and to answer question "Why is it happening?" Predictive analytics try to figure out the future trends and scenarios, addressing the future output parameters and answering question "What will be happening?" Prescriptive analytics going further trying to figure out the optimisation of input parameter in order to achieve the expected output parameters.

Predictive and prescriptive analytics rely on intense calculations and thus require higher capacity in terms of computation resource. Nevertheless, all types of analytics aim to process the largest dataset as quickly as possible in order to improve accuracy of results and that requires capacity in terms of data storage.

Advanced Analytics Advanced Analytics are comprised with number of practices and technologies, including data mining, predictive analytics, natural language processing, and artificial intelligence such as machine learning, decision trees, and neural networks. Advanced Analytics involve statistical, quantitative, or mathematical analysis of data and developing testing, training, scoring and monitoring models. [6]

Advanced Analytics are used to discover why something happens, what happens next, and how to optimise actions to achieve desired results. Advanced Analytics mostly need to explore raw, detailed data rather than small samples and aggregations (designed for BI and OLAP). [6]

Advanced Analytics, sometimes called Explanatory Analytics, offer highly complementary technologies to the Business Intelligence. Analytic models (with phases training, deployment, monitoring) are running against data or event streams and requiring the computation power.

High Performance High Performance is important in phase of a model deployment. Demands on real-time decisions in operations supported by analytics increase massively data movements and replications. High performance methods can be considered massively parallel processing, grid computing, in-database, in-memory, complex event processing, stream processing, etc.

5. CONCLUSION

Since the Big Data has been continuously identified for a decade there is a lot research done in literature, white papers, and online references. In this thesis, the Big Data Phenomenon is summarised in an overview including its causalities, definition, influence and impacts. It represents a starting point and driver for High Performance Analytics in terms of raw material that contains hidden information, patterns and value. Concluding from research, Big Data, with its dynamic dimensions, should not be considered as a problem, rather than opportunity to turn it into advantage.

High Performance Analytics is extensively researched in this thesis as an approach towards handling Big Data. Due to this area it is still emerging, being refined and formalised among vendors, research on HPA is challenging in order to bring overview, classification of HPA methods and techniques (In-Memory Analytics, In-Database Analytics, and Parallel Computing), their characteristics, and appropriate usage.

HPA is driven by business world with broad requirements to compute results as fast as possible on the largest dataset. The formation HPA becomes possible with technological evolution (Very Large Memory, 64bit address, Grid Computing) and affordability of hardware (costs, price:performance indicator). For now, HPA can be seen as a solution complementary to the Business Intelligence, but highly on premises the evolution will continue further. The research could be extended to dive into HPA solutions from another vendors comparing multiple proprietary approaches in details.

REFERENCES

[1]. High-Performance Analytics; SAS Institute - White paper; 2012.

[2]. Big Data Analytics: Future architectures, Skills and roadmaps for the CIO; Philip Carter; IDC White Paper; 2011.

[3]. From Big Data to Meaningful Information - Insights from a webinar sponsored by KMWorld Magazine and SAS; Conclusion paper; 2013.

[4]. Big Data Meets Big Data Analytics; SAS Institute - White paper; 2012.

[5]. Data Visualisation Techniques; SAS Institute - White paper; 2012.

[6]. Big Data Analytics: Profiling the Use of Analytical Platforms in User Organizations; Wayne Exkerson; 2011.

[7]. Seven Keys to High-Performance Data Management for Advanced Analytics; TWDI Monograph Series; David Stodder; 2011.

[8]. Big Data Analytics; TDWI best practices – Fourth Quarter 2011; Report Philip Russom; 2011.

[9]. Intelligence Quarterly, Second Quarter 2012; SAS Institute; 2012.

[10]. Postmoderní architektura Business Intelligence; Business Intelligence Fórum 2012; Vladimír Kyjonka; 2012.

[11]. V koži pilota high-end stíhačky; ITnews; Vladimír Kyjonka; 2013; online source: http://www.itnews.sk/2013-02-04/c153968-v-kozi-pilota-high-end-stihacky

[12]. What is Big Data; IBM; 2012; online source: www.ibm.com/software/data/bigdata/

[13]. CRM Data Strategies: The Critical Role of Quality Customer Information; Gartner Inc.; 2003.

[14]. Bitpipe research guide: Business Intelligence Overview; online source: http://www.bitpipe.com/bi/bi_overview.jsp

Author Details:

S. Harini Krishna working as an Assistant professor in department of CSE at Geethanjali college of Engineering and Technology affiliated to JNTU Hyderabad. She has 6+ years teaching experience. Her Research interests include Cloud Computing and Artificial Intelligence.

K. Gnana Mayuri working as an Assistant professor in department of CSE at Geethanjali college of Engineering and Technology affiliated to JNTU Hyderabad. She has 6+ years teaching experience. Her Research interest includes Big data, Information Security and Cloud Computing.