Challenges of the Fourth Paradigm

S. Ramakrishnan
B.Sc. M.C.A
Research Scholar
Alagappa University, Karaikudi, Tamilnadu, India

V. Nanda Kumar. Ph.D
Programmer Selection Grade, Computer Centre,
School of Computer Science Alagappa University,
Karaikudi, Tamilnadu, India

ABSTRACT
Even a decade ago, data storage scalability was a major technical issue. Efficient and scalable technology with data management and storage has almost eliminated the issue. Cloud computing has reduced the initial IT investments for all businesses. Large amounts of data are generated from the internet and by organizations. Big Data is new paradigm also termed the Fourth paradigm. The term "Big Data" envelops several factors of IT ranging from technology to economic and business models. The basic concept is in providing anytime resources to all with adaptability to known and new service demands. This paper discusses the challenges rising from the use of Big Data and suggestions to overcome the challenges.

General Terms
Big Data, Privacy in Big Data

Keywords
Fourth Paradigm, Stream Processing, Big Data Analytics

1. INTRODUCTION
Big Data is the use of techniques to capture, process, analyze and visualize large datasets. The technologies involved in Big Data are called Big Data technologies. Big Data aims at any-time availability of services and resources. Big Data is an innovation which is productive [1] and offers new business opportunities with challenges. Re-usage of public data can give consumers more choice and more value for their money and at the same time generate new businesses and jobs. The European Commission has initiated an Open Data strategy by amending Public Sector Information Directives and encouraging openness and reuse of public data [2]. Big Data is important to scientific infrastructures and development of innovative value added services. It is also important to other research areas which are experiencing the need for efficient usage [3]. Content management is another applicable area for Big Data. The challenges associated with the use of Big Data are many. Data generated from organizations data makes Big Data the largest source and a base for success of social platforms [1]. Google and Amazon who initially faced increasing data volumes designed ad-hoc solutions to handle them and formed the base for the Big Data Trend when people started these large sources of data. Ability to process this kind of data in many dimensions became a new found opportunity for traditional IT vendors and Small businesses. Many vendors are working on improvements in this emerging and important field.

2. SOCIO-ECONOMIC INFLUENCES
Studies reveal, developed countries have the biggest potential and impact through the use of Big Data [4] as Big Data is expected to add more revenues [1]. Countries could benefit from the cumulative financial and social impact of Big Data. Big Data analytics is impacting organizations as they produce streams of real time data gathered which can be integrated in multiple application domains. Big Data is an entirely new dimension to social interactions. Moreover, Experiments in science generate vast volumes of experimental data where Petabytes (PB) of data per day is not uncommon. Scientific data requires proper handling and solutions to analyze huge datasets and help discoveries. Mobile phone traffic is expected to grow to 10 Exabyte per month by 2016, mainly due to the increased usage of smart phones and tablets [5]. Big Data technology is needed in order to realize some advanced use cases in today's mobile networks and will be certainly be required in future networks. Big Data can be used to gain insights into network quality and identify faults, support security and management. The web media with new types of content called New Media are generating huge amounts of contents. New media participants report, blog and tweet away from conventional journalism which has been broadcasting and mediating mainstream contents from authoritative sources. The mainstream journalism is being disrupted due to the internet web technologies. Big Data can be a start to scale legacy systems like Online Analytical Programming (OLAP) systems which perform analysis on online data.

3. CHALLENGES
Big Data challenges emanate mainly from techniques that need to use current technology's capacity to store and query available data sets, instead of sampling. This availability has tremendous implications in research areas starting from machine learning to classification. Moreover, the data mining techniques on big Data need a solid scientific foundation in design. New efficient, scalable and implementable algorithms are another challenging task. The data structures also need to be understandable and scalable for these techniques. One of the main obstacle in analytics is the lack of understanding of business data structures for improvement [6]. The new objects need modeling and simulation which are complex and massive in nature with vast distributions. Strong computational paradigms orienting towards distributed and parallel computing, data visualization and simulation are the need for the day.

3.1 Data Context
Data context in Big data is an imminent part due to the increasing number of users and devices. Efficient content-aware routing of data is required since, Big Data solutions focus on processing and routing the entire data immediately. Filtering data that is not related to a context is a challenging task. For example manufacturing data has very little

importance to the user´s personal information and needs to be filtered.

Suggestion 1: Combining contexts in Big Data can be an attractive paradigm to query and process heterogeneous data streams and improve the quality of a mining process. Context-aware Big Data solutions could exploit and focus on portions of data in the first approximation to maintain a high probability of hit for all application relevant events. Awareness of the data can optimize management of resources, systems and services in many application domains with manifest advantages in terms of cost reduction and complexity decrease. Context awareness is generally useful in reducing resource consumption. For instance switching off sensors can decrease duty cycles which are expensive in terms of power consumption and traffic bandwidth. Industry-oriented research along this guideline is needed for complexity reduction of Big Data solutions.

3.2 Imagining Data

The structure of Data is vital for effective utilization of Big Data, since analytics reports may be in different formats. Pure textual representation requires careful presentations due to their unstructured nature while complex information presentation, with its associated tasks, increases the design issues multifold. Further, the information interface needs to be responsive to human needs and knowledge. Researchers need to understand the relevant information, since, too much of information cannot be searched or researched efficiently. Researchers need to understand the relevance and relatedness of information lack of which can lead to a degradation in performances.

Suggestion 2: Researching industrial applicability to Big Data to achieve the probability of hits is highly important and still necessitate significant investigation. The relevance and relatedness of information should be promoted to researchers and application developers by the respective domain experts. Develop corresponding analysis tools with supporting user interaction techniques enables easy transitions from one scale or form of aggregation to another. It is necessary to vividly understand the data as a naïve user understands for maximum efficiency.

3.3 Scalable Data management

The natural expansion of internet has created an increase in the global data production. Performance and scalability challenges in Big Data solutions and techniques are technical issues as they deal with the huge volumes of stored data. Novel and effective solutions dealing with data volumes are a challenging task as they need to perform and be scalable. Also Big Data analysis has to be performed within time constraints as defined by respective application domains for cost reductions. Focus on process mining is essential for progression of Big Data management.

Suggestion 3: Defining quality constraints on both Big Data storage like replication levels and processing like parallelism on computing resources should be considered. Integrating unstructured data with structured models should be considered for total solutions. New Big Data-specific parallelization techniques and automated distribution of tasks should be used for effective real-time processing. New frameworks and open APIs for the quality-aware development efforts should be in place for application developers and domain experts.

3.4 Stream Analytics

One of the most open Big Data technical challenge is the storage, management and processing of voluminous data streams. Business decisions do not desire long waiting hours for results of analytic processes and expect a near real-time response from systems. Handling large amounts of streaming data, ranging from structured to unstructured is challenging in Big Data because the data is heterogeneous, voluminous and highly dynamic. For example data collection related to a disaster in an area can easily occupy terabytes in data streams may have to accommodate gigabytes of data in minutes. The capabilities of existing systems to process streaming information and answer queries are limited. New approaches are needed for developing tools and techniques for querying continuous streams.

Suggestion 4: Open APIs and opening the field to small/medium-size companies can help reducing the entrance barriers to this potential market. Reduction of the initial investments needed for new stream processing-based applications. Leveraging related developers/users communities can help building business ecosystem with a clear understanding of legal and regulatory issues between countries, cross-enterprise collaboration issues, the nature of data sources and translations to be applied on them**.**

3.5 Distributed Sets

Every Organization's data requires analysis and Big Data analytics use **D**istributed data sets of organizations. Any business data when expanded to related domains increases the information structures and culminates in distributed information sets. Performing meaningful analytics on these volumes of data becomes a major challenge. Though Cloud based solutions have lowered storage costs and helped transition from a managed infrastructure to a service based infrastructure, these large volumes of data need to be distributed to share the workload.

Suggestion 5: A positive solution to this situation is virtualization of storage in data centre's with a generalization of the cloud based solutions. The noSQL is an answer to store and query huge volumes of heavily distributed data.

3.6 Data Validation

Validating information in Big data is again a major challenge, since the information sources are heterogeneous and varied. Each data source has a different form like blogs, tweets, articles, comments etc. The complexity lies in the fact that human expressions in a language differ and a standard level of computational validation becomes less feasible.

Suggestion 6: Derivation of simple rules with user specific validations on content with leverage on recommendations from users, can be a possible solution. The recommending users can assessed on the basis of domain knowledge, trust and reputation. Further, learning algorithms need to update rules based on user feedback. Machine learning algorithms can be an efficient solution for extracting patterns or rules in Big Data, semi-automatically or automatically**.**

3.7 Privacy

Collection, storage and analysis of personal data has dramatically increased with big data. Many companies built their business models on using and selling user profiles generated from these data sources. Governments analyze information exchange between citizens. User information can

be extracted for surveillance and marketing purposes from existing internet logs. Tools track identity of users, cameras are used in surveillance, mobile phones send location information, debit and credit cards display amounts spent. Social media allows user-to-user contact and access to pictures, videos and movies. Data protection and information security are particularly sensitive issues and querying them without fuss is a tricky issue. For Example Medical data. The security and legal issues on these distributed data sets presents one of the most complex problems to solve. The slightest doubt that health records could fall into the wrong hands freezes Big Data adoption.

Suggestion 7: One promising approach for preserving privacy is the use of data marts. Big Data Analytics Agents (BDAA) performing a specific function or set of functions on data sets and dispatching them to appropriate locations can be security controlled by the receiving data sets. BDAA's could be designed for the specific features of Big Data like analyzing video for predefined features.

4. CONCLUSION

Big Data is the emerging and future paradigm of ICT called the Fourth paradigm. To be able to extract the benefits of Big Data, it is crucial to use intelligently, manage re-use of Data Sources including public and build useful applications and services. The key points stressed in this paper seek to ensure that the necessary technical conditions, opportunities in big data suggestion to regain a competitive position in Big Data Technologies. With the rapid growth on information being displayed it is essential to arrange the data in an easy and accessible way to search a particular subject. New designs on Big Data for databases and efficient ways to support massively parallel processing have led to a new generation of products like the so called noSQL databases and the Hadoop map-reduce platform. Big Data business ecosystem can only be built with a clear understanding of policies addressing legal and regulatory issues between countries and within the context of a revised Data protection legal framework. Promising areas that call for further investigation and industrially applicable results include effective non-uniform replication, selective multi-level caching, advanced techniques for distributed indexing, and distributed parallel processing over data subsets with consistent merging of partial results.

5. REFERENCES

[1] Big Data: The next frontier for innovation, competition and productivity, June 2011, McKinsey Global Institute,

[2] "Big Data at your Service", http://ec.europa.eu/information_society/newsroom/cf/dae/itemdetail.cfm?item_id=8337

[3] Neelie Kroes, http://www.europarl.europa.eu/sides/getAllAnswers.do?reference=E-2012-006126&language=EN

[4] Data Equity – Unlocking the value of big data, p. 4 ff. April 2012, Center for Economics and Business Research

[5] Mobile is the new face of engagement, Forrester , February 2012

[6] Big Data, Analytics and the Path From Insights to Value, Dec 2010, LaValle et alTavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.