Volume 3, Issue 8, August 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Study and Comparison of Big Data with Relational Approach

Payal Malik[1] , Lipika Bose[2]

[1]Master in Technology,
Amity University,
Noida, Uttar Pradesh, India

Abstract— Innovations in technology and greater affordability of digital devices have presided over today's Age of Big Data, an umbrella like term for the sudden increase in the quantity and diversity of high frequency digital data. "Big Data" is the latest buzz word in the technical space. Do you wonder "What big Data is? Why is it called big?"

These data hold the potential—as yet largely unused—to allow decision makers to track development progress, improve social protection, and understand where existing policies and programmers' require adjustment. Big Data has the potential to revolutionize not just research, but also education. In this paper, we discuss several research activities. This data could be used to design the most effective approaches to education, starting from reading, writing, and math, to advanced, college-level, courses. This paper does not offer a impressive theory of technology-driven social change in the Big Data era. Rather it aims to outline the main concerns and challenges raised by "Big Data for Development".

Keywords— Big Data, innovation, revolutionize, social protection, challenges, technology-driven social change.

I. INTRODUCTION

Big Data refers to datasets that grow so large that it is difficult to capture, store, manage, share, analyze and visualize with the typical database software tools. Big data is considered at the heart of modern science and business. Big data is a term which is applied to a new generation of software, applications, system and storage architecture, which are all designed to derive business value from unstructured data. As the name suggests that it is "Big". But how big, you cannot imagine until this point. All of us are very much familiar with GB-Giga Bytes, few know about TB-Tera Bytes too (Both are unit to measure the computer memory size). Very few know about PB (Penta Bytes) and EB (Exa Bytes). 1 EB is equivalent to 106 TBs or 1 billion of GBs. The unit is huge and it tells the vastness of its name. The size4 of Big Data starts at the point, where a common man's thinking stops. Each day, close to 3 EBs of data is being generated and one of the leading Market Research firm has announced that 1200 EB of data will be generated in the year 2012 only.

Who is creating this data? The answer is "We – The internet freaks".

You wonder "How?" Whenever you write an email or a blog or a sweet little tweet or even shorter "Like" on FB, organization capturing web traffic, Analytics, IT system logs, Music. Videos, Games, Website contents, eBooks, Government docs, corporate data etc are name to few. Collectively this whole of the data is called "BIG DATA".

Generating the huge data is just the start. A whole new world of technology starts when this data needs to be analyzed and bring out some logical information. The major internet firms like Google, Amazon, and eBay etc first realize that the data being generated over the net is expanding at a much greater pace. This data needs to be tamed from the perspective of Economics. This data need to be used in the corporation's favour and interest.

What exactly is the potential applicability of "Big Data for Development?" At the most general level, properly analysed, these new data can provide snapshots of the well-being of populations at high frequency, high degrees of granularity, and from a wide range of angles, narrowing both time and knowledge gaps. Practically, analyzing this data may help discover what Global Pulse has called "digital smoke signals" [1]. Value is created in the big data design segment by using data to drive product innovation, improve time-to-design, create transparency in the process flow, and by massively reducing the cost of physical design (or by replacing it all-together). Like For example, 3D modeling software allows aeronautic engineers and automotive designers to experiment with esoteric design without the costly need for physical models or wind tunnels. Researchers are in the both big data creators as well as big data users.

II. KEY ELEMENTS OF BIG DATA

Big Data has the following key elements or characteristics:

A. Volume

Even though the early pioneers of Big Data have been the largest, web based, social media companies Google, Yahoo, Face book – it was the volume of data generated by their services that required a radically new solutions rather than the need to analyze social feeds.

B. Variety

Refers to the many different data and file types that are important to manage and analyze more thoroughly, but for which traditional relational databases are poorly suited. Some examples of this variety include sound and movie files, images, documents, geo-location data, web logs, text strings, web contents etc. Big data is probably better termed "multi-structured" as it could include text strings, documents of all types, audio and video files, metadata, web pages, email messages, social media feeds, form data, and so on.

C. Velocity

It is about the rate of change in the data and how quickly it must be used to create real value. Traditional technologies like Ab Initio, ETL, Data Base etc are especially poorly suited to storing and using high- Velocity data.

D. Viscosity

It measures the resistance to flow in the volume of data. This resistance can come from different data sources, friction from integration flow rates, and processing required turning the data into insight. Technologies to deal with the viscosity include improved streaming, agile integration bus, and complex event processing.

E. Virality

It describes how quickly information gets dispersed across people to people (P2P) networks. Virality measures how quickly data is read, spread and shared to each unique node. Time is a determinant factor along with the rate of spread. Example is as it happens during a 'like' on face book. You like one image that is shown to your friend's timeline, they also see and like and like and the word is spread.

III. WHAT MAKES BIG DATA DIFFERENT? COMPARING INFORMATION ARCHITECTURE OPERATIONAL PARADIGMS

Big data differs from other data realms in many dimensions. In the following table you can compare and contrast the characteristics of big data alongside the other data realms described in Oracle's Information Architecture Framework (OIAF).

Table 1: Data realms described

Data Realm	Structure	Volume	Description	Examples
Master Data	Structured	Low	Enterprise-level data entities that are of strategic value to an organization. Typically non-volatile and non-transactional in nature	Customer, product, supplier, and location/site
Transaction Data	Structured & semi-structured	Medium – high	Business transactions that are captured during business operations and processes	Purchase records, inquiries, and payments
Reference Data		Structured & semi-structured	Low – Medium	Internally managed or externally sourced facts to support an organization's ability to effectively process transactions, manage master data, and provide decision support capabilities.
Metadata		Structured	Low	Defined as "data about the data." Used as an abstraction layer for standardized descriptions and operations. E.g. integration, intelligence, services.
Analytical Data	Structured	Medium-High	Derivations of the business operation and transaction data used to satisfy reporting and analytical needs.	Data that reside in data warehouses, data marts, and other decision support applications.

Documents and Content	Unstructured	Medium – High	Documents, digital images, geo-spatial data, and multi-media files.	Claim forms, medical images, maps, video files.
Big Data	Structured, semi-structured, & unstructured	High	Large datasets that are challenging to store, search, share, visualize, and analyze.	User and machine-generated content through social media, web and software logs, cameras, information-sensing mobile devices, aerial sensory technologies, and genomics.

These different characteristics have influenced how we capture, store, process, retrieve, and secure our information architectures. As we evolve into Big Data, you can minimize your architecture risk by finding synergies across your investments allowing you to leverage your specialized organizations and their skills, equipment, standards, and governance processes.

Table 2: Data Realm Characteristics

Data Realms	Security	Storage & Retrieval	Modeling	Processing & Integration	Consumption
Master data Transactions Analytical data Metadata	Database, app, & user access	RDBMS / SQL	Pre-defined relational or dimensional modeling	ETL/ELT, CDC, Replication, Message	BI & Statistical Tools, Operational Applications
Reference data	Platform security	XML / xQuery	Flexible & Extensible	ETL/ELT, Message	System-based data consumption
Documents and Content	File system based	File System / Search	Free Form	OS-level file movement	Content Mgmt
Big Data - Weblogs - Sensors - Social Media	File system & database	Distributed FS / noSQL	Flexible (Key Value)	Hadoop, MapReduce, ETL/ELT, Message	BI & Statistical Tools
Data Realms	Security	Storage & Retrieval	Modeling	Processing & Integration	Consumption
Master data Transactions Analytical data Metadata	Database, app, & user access	RDBMS / SQL	Pre-defined relational or dimensional modeling	ETL/ELT, CDC, Replication, Message	BI & Statistical Tools, Operational Applications
Reference data	Platform security	XML / xQuery	Flexible & Extensible	ETL/ELT, Message	System-based data consumption

IV. BIG DATA ARCHITECTURE

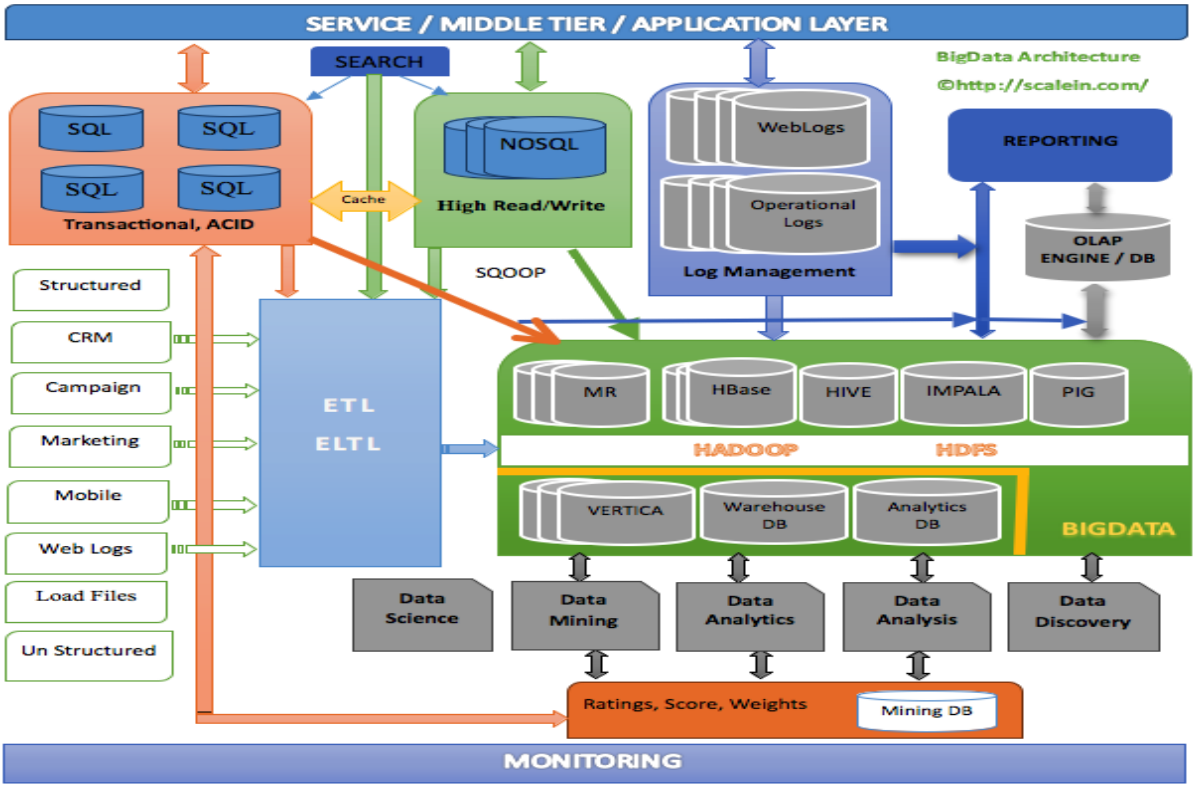
This is the typical "Big" data architecture, which covers most components involved in the data pipeline. More or less, we have the same architecture in production in number of places (with some varying components due to data sources and data consumption is varied from company to company).

Any data architecture loosely consists of four major logical components:

A. Data Source:

True source of data are coming from heterogeneous data sources. This is typically your data stores (SQL or NoSQL) that give a structured data or any other data coming through APIs or other means (semi-structured or un-structured).

Figure 1.Bigdata architecture



B. Data Transformation

Transformation of data from one form to another, its either part of **ETL** (**E**xtract, **T**ransform and **L**oad) or import/export tools and/or scripts. Mainly used to load all sources of data into data processing pipeline.

Log management tools can also be considered as part of ETL, as they generate useful events from log files and present dashboard with alerting system in place or they can be directly loaded.

C. Data Processing or Data Integration

This is yet another source of vast data by combining both structured and un-structured data in one place (either real-time or incremental loading); mainly for data processing (Data Warehousing or Analytics) and generates usable data (materialized or aggregated) that can be consumed by data consumption components.

- Hadoop and Ecosystem (Hadoop/HDFS, Map-reduce, HBase, Hive, Impala, Pig etc) – uses HDFS as native storage
- Data Warehouse and Analytics solution (MySQL, SQL Server, Vertica, Green Plum, Aster Data, Exadata, SAP HANA, IBM Netezza, IBM Pure Data, Tera Data, etc.) – Uses vendor specific storage, optionally uses HDFS, even though with degraded performance.
- In-memory Analytics (SAS, Kognitio, Druid, etc.). This is an emerging market and trying to take advantage by reading directly from HDFS. We will see lot of in-memory analytics in coming days.

D. Data Consumption

Data consumption components that either consumes or exposes the data in usable form to end users or to other layers internally (ad-hoc) or externally (using APIs)

Reporting (custom dashboards, micro strategy, pentaho, business objects, cognos, Hyperion, tableau, etc.) Search or Discovery (solr, elastic search, tibco spotfire, datameer etc.)

Data Science, Mining and Analysis (mainly for internal data analysis to predict or estimate the overall performance and also drive recommendation using set of algorithms, user defined map-reduce jobs or ad-hoc queries)

Apart from the four logical components, **monitoring** plays a crucial role in detecting any failure within the data pipeline along with threshold changes to identify any bottlenecks in terms of performance, scalability and overall throughput.

No SQL Vs SQL

Table 3 NOSQL vs SQL

	SQL Databases	NoSQL Databases
Types	One type (SQL database) with minor variations	Many different types including key-value stores, document databases wide-column stores, and graph databases
Development History	Developed in 1970s to deal with first wave of data storage applications	Developed in 2000s to deal with limitations of SQL databases, particularly concerning scale, replication and unstructured data storage
Examples	MySQL, Postgres, Oracle Database	MongoDB, Cassandra, HBase, Neo4j
Data Storage Model	Individual records (e.g., "employees") are stored as rows in tables, with each column storing a specific piece of data about that record (e.g., "manager," "date hired," etc.), much like a spreadsheet. Separate data types are stored in separate tables, and then joined together when more complex queries are executed. For example, "offices" might be stored in one table, and "employees" in another. When a user wants to find the work address of an employee, the database engine joins the "employee" and "office" tables together to get all the information necessary.	Varies based on NoSQL database type. For example, key-value stores function similarly to SQL databases, but have only two columns ("key" and "value"), with more complex information sometimes stored within the "value" columns. Document databases do away with the table-and-row model altogether, storing all relevant data together in single "document" in JSON, XML, or another format, which can nest values hierarchically.
Schemas	Structure and data types are fixed in advance. To store information about a new data item, the entire database must be altered, during which time the database must be taken offline.	Typically dynamic. Records can add new information on the fly, and unlike SQL table rows, dissimilar data can be stored together as necessary. For some databases (e.g., wide-column stores), it is somewhat more challenging to add new fields dynamically.
Scaling	Vertically, meaning a single server must be made increasingly powerful in order to deal with increased demand. It is possible to spread SQL databases over many servers, but significant additional engineering is generally required.	Horizontally, meaning that to add capacity, a database administrator can simply add more commodity servers or cloud instances. The NoSQL database automatically spreads data across servers as necessary
Development Model	Mix of open-source (e.g., Postgres, MySQL) and closed source (e.g., Oracle Database)	Open-source
Supports Transactions	Yes, updates can be configured to complete entirely or not at all	In certain circumstances and at certain levels (e.g., document level vs. database level)
Data Manipulation	Specific language using Select, Insert, and Update statements, e.g. SELECT fields FROM table WHERE…	Through object-oriented APIs

	SQL Databases	NoSQL Databases
Consistency	Can be configured for strong consistency	Depends on product. Some provide strong consistency (e.g., MongoDB) whereas others offer eventual consistency (e.g., Cassandra)

V. CHALLENGES AND OPPORTUNITIES WITH BIG DATA

A. Challenges In Big Data

We have some common challenges that underlying in big data they are:-

1) *Heterogeneity and Incompleteness:* Machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step in (or prior to) data analysis. Even after data cleaning and error correction, some incompleteness and some errors in data are likely to remain. This incompleteness and these errors must be managed during data analysis. Doing this correctly is a challenge.
2) *Scale:* Of course, the first thing anyone thinks of with Big Data is its size. After all, the word "big" is there in the very name. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades.
3) Timeliness: The flip side of size is speed. The larger the data set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when one speaks of Velocity in the context of Big Data.
4) *Privacy:* The privacy of data is another huge concern, and one that increases in the context of Big Data.
5) *Human Collaboration*: In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding. In today's complex world, it often takes multiple experts from different domains to really understand what is going on. A Big Data analysis system must support input from multiple human experts, and shared exploration of results.

B. Opportunities In Big Data

1) *Algos:* Big Data initiatives are driving increased demand for algorithms to process data, as well as emphasizing challenges around data security and control, and minimizing impact on the existing system. Robust and fast algorithms would be the need of the hour. Companies like goggle, face book, twitter are exploiting the power of algorithms for their benefits. As a professional, learning about the power of algorithms will be adding a boost.
2) *Mobility:* Mobile applications and internet-connected devices such as tablets and smart phone are creating greater pressure on the ability of technology infrastructures and networks to consume, index and integrate structures and unstructured data from a variety of sources. All the data being carried through mobile devices is being captured and analyzed.
3) *Extract, Transform, Load:* Population of centralized data warehouse systems will require traditional ETL processes to be re-engineered with big data frameworks to handle growing volumes of information. This will create the need of Data Scientists.
4) *Unlocking the Value of Data:* Advances in Big Data storage and processing frameworks will help financial services unlock the value of data in their operations departments in order to help reduce the cost of doing business and discover new arbitrage opportunities. This will create the need of thinkers/ managers, who can provide the ideas to monetize the value of data.
5) *Enterprise Risk Management:* Financial institutions are ramping up their enterprise risk management frameworks, which rely on the master data management strategies to help improve enterprise transparency, audit and executive oversight of risk.
6) *Governance and Risk reporting:* New regulatory and compliance requirements are placing greater emphasis on governance and risk reporting, driving the need for deeper and more transparent analysis across global organizations.

VI. CONCLUSION

The Big Data has got the tremendous opportunities in coming years. The firms like facebook, twitter, Zynga, Google, TCS, IBM, Infosys etc are high on concept and are working to tame the same. As a professional – Students, Academia, II Professionals, Hardware companies have a lot to learn and perform. There are lot of opportunities for emerging markets like Brazil, India and China etc. The channels are increasing and hence the data is increasing. The predicate analysis is becoming easier and strong. According to some experts, "by employing massive data mining, science can be pushed towards a new methodological paradigm which will go beyond the boundaries between theory and experiment." Another point of view

frames this new ability to disclose some facts from large datasets as "the fourth paradigm of science". SQL is an extremely powerful way to manipulate and understand data. Until now, Hadoop's SQL capability has been limited and impractical for many users. HAWQ is the new benchmark for SQL on Hadoop — the most functionally rich, most mature, most robust SQL offering available.

References

[1] Helen Sun, Heller Peter An Oracle White Paper in Enterprise Architecture August 2012 < www.oracle.com/bigdata.>
[2] Kirkpatrick, Robert. "Digital Smoke Signals." UN Global Pulse. 21 Apr. 2011. http://www.unglobalpulse.org/blog/digital-smoke-signals.
[3] "The Data Deluge." The Economist. 25 Feb 2010. <http://www.economist.com/node/15579717> and Ammirati, Sean. "Infographic: Data Deluge – 8 Zettabytes of Data by 2015." Read Write Enterprise. http://www.readwriteweb.com/enterprise/2011/11/infographic-data-deluge---8-ze.php
[4] King, Gary. "Ensuring the Data-Rich Future of Social Science." Science Mag 331 (2011) 719-721. 11 Feb, 2011 Web http://gking.harvard.edu/sites/scholar.iq.harvard.edu/files/gking/files/datarich_0.pdf
[5] Helbing, Dirk , and Stefano Balietti. "From Social Data Mining to Forecasting Socio-Economic Crises." Arxiv (2011) 1-66. 26Jul2011 http://arxiv.org/pdf/1012.0178v5.pdf.
[6] Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh and Angela H. Byers. "Big data: The next frontier for innovation, competition, and productivity." McKinsey Global Institute (2011): 1-137. May 2011.
[7] " World Population Prospects, the 2010 Revision." United Nations Development Programme. http://esa.un.org/unpd/wpp/unpp/panel_population.htm
[8] Gray, Jim (ed. Gray, J., Tansley, S. and Tolle, K.). "eScience: A transformed scientific method." The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research. Redmond, Washington, 2009. < http://research.microsoft.com/en- s/collaboration/fourthparadigm/contents.aspx>.
[9] http://unglobalpulse.org/
[10] McKinsey Global Institute
[11] Challenges and Opportunities with Big Data A community white paper developed by leading researchers across the United States
[12] http://hadoop.apache.org/

