Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Science and Engineering

Elixir Comp. Sci. & Engg. 69 (2014) 23461-23465

ISSN: 2229-712X

Big data: Magnification beyond the relational database and data mining exigency of cloud computing

Abhishek Khare

Department of Computer Science and Applications, Shri Vaishnav Institute of Management Indore, India.

ARTICLE INFO

Article history:

Received: 31 October 2013;
Received in revised form:
20 April 2014;
Accepted: 23 April 2014;

Keywords

Data Mining;
Relational Database;
Big Data;
Cloud Computing;

ABSTRACT

Today the term big data draws a lot of attention. Big data is the new technology that enables an organization to take advantage of the natural resource of big data. Organizations store up huge amounts of information, they can mine those databases to discover subtle patterns, correlations, or relationships that human brains can't perceive on their own because the scales involved are beyond our ability to process. Big data is a hot topic in the market today and attracted a lot of attention from retail, automotive, and manufacturing industry, government as well as academia. This paper introduces several enhancements in big data over data mining and relational database. Finally, I discuss the urgency of big data processing in cloud computing environments.

Introduction

Big Data, loosely defined, is the ability to gather, analyze, interpret and most importantly act on large volumes of data to identify and solve problems. Big Data is not the Created Content nor is it even its Consumption- it's the analysis of all the data surrounding or swirling around it [1]. IDC defines-Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis. We are experiencing a data revolution in both sciences and industry. This is in part the result of cheaper data capture and storage technology, but it is also the culmination of a cultural transformation in the commercial world, whereby data are now viewed as a source of actionable insights, much like in the sciences. Largely the focus of the Big Data revolution has until recently been on infrastructural developments, such as the Apache Hadoop project. And yet, the availability of Big Data can be little more than a lost opportunity without respectively powerful analytics. The statistics and machine learning communities have risen to this challenge, with a growing part of their respective literature dedicated to the challenges posed by Big Data. For decades, companies have been making business decisions based on transactional data stored in relational databases. Beyond that critical data, however, is a potential treasure trove of non-traditional, less structured data: weblogs, social media, email, sensors, networks data and photographs that can be mined for useful information. Decreases in the cost of both storage and compute power have made it feasible to collect this data - which would have been thrown away only a few years ago. As a result, more and more companies are looking to include non-traditional yet potentially very valuable data with their traditional enterprise data in their business intelligence analysis [3].

The economic, industrial, commercial, social, political and sustainability problems we face cannot be successfully addressed using the management techniques and models largely inherited from the Industrial Revolution. The world no longer appears infinite in resources, slow paced, linear and stable. We now see the limitations; feel the impact of rapid change; and we can conceptualize the non-linear and unstable nature of it all! We are also starting to comprehend the scale and the need for machine assistance.

Sophisticated computer models for weather systems are now complemented by ecological, economic, conflict and resource modeling of varying depth and accuracy. However, the key is always the accuracy and coverage of the primary data. We started with modest databases and data mining, but they mostly proved inadequate, and we are now amassing vast databases on every aspect of life - people, planet and machines. This 'BIG DATA' explosion demands a rethink of how, what, and where we gather data; the way we analyze and model; and the way we make decisions.

Data Mining Challenges and Role of Big Data

Data mining is a powerful new technology with great potential to help companies focus on the most important information in the data they have collected about the behavior of their customers and potential customers. It discovers information within the data that queries and reports can't effectively reveal. This paper emphasizes many aspects and challenges of data mining and how efficient is big data.

Data mining is a method of pattern discovery against a pool of data using specialized data mining tools. These tools use a sophisticated blend of classical and advanced components like artificial intelligence, pattern recognition, databases, traditional statistics, and graphics to present hidden relationships and patterns they find in any given data pool. One of the official definitions for data mining is: "Data analysis without preconceived hypothesis to unearth unsuspected or unknown relationships, patterns or associations of data." Simply put, "without preconceived hypothesis" means you don't know what exactly you are looking for, "to unearth" means the tool will analyze the data using special algorithms and analytical models to discover any patterns in the data and then tell you about them. The term data mining is sometimes misused to mean "ability to write a lot of different SQL queries."

Tele:
E-mail addresses: abhishekkhare@rocketmail.com

Data Mining was limited, planer, linear and constrained to a few relationships amongst people: what they did, where they went, who they knew and so on. In contrast; Big Data is unbounded, spans all peoples and machines in all domains and activities with application to every aspect of life, business, industry, government and sustainability etc. It also takes into account the non-linear nature of relationships and events

Big Data is an almost unconscious outcome of the desire and need to sustain all peoples on a rapidly smaller looking planet [2].

Major research challenges of data mining

In this section, we will examine several major challenges raised in science and engineering from the data mining perspective, and point out some promising research directions and importance of Big Data

Information network analysis

With the development of Google and other effective web search engines, information network analysis has become an important research frontier, with broad applications, such as social network analysis, web community discovery, terrorist network mining, computer network analysis, and network intrusion detection. However, information network research should go beyond explicitly formed, homogeneous networks (e.g. web page links, computer networks, and terrorist e-connection networks) and delve deeply into implicitly formed, heterogeneous, and multidimensional information networks. Science and engineering provide us with rich opportunities on exploration of networks in this direction [4].

There are a lot of massive natural, technical, social, and information networks in science and engineering applications, such as gene, protein, and microarray networks in biology; highway transportation networks in civil engineering; topic- or theme-author-publication-citation networks in library science; and wireless telecommunication networks among commanders, soldiers and supply lines in a battle field. In such information networks, each node or link in a network contains valuable, multidimensional information, such as textual contents, geographic information, traffic flow, and other properties. Moreover, such networks could be highly dynamic, evolving, and inter-dependent.

Big Data, loosely defined, is the ability to gather, analyze, interpret and most importantly act on large volumes of data to identify and solve problems. Hospitals use it to prevent illness. The financial industry uses it to detect credit card fraud. Airlines use it to fill seats. And, Amazon uses it to tell you what you might like to read next. So what's Big Data have to do with library discovery and the Summon service? The ability to leverage Big Data is making it possible to better understand how users perform research.

Discovery, understanding, and usage of patterns and knowledge

The Scientific and engineering applications often handle massive data of high dimensionality. The goal of pattern mining is to fiend item sets, subsequences, or substructures that appear in a data set with frequency no less than a user-specified threshold. Pattern analysis can be a valuable tool for finding correlations, clusters, classification models, sequential and structural patterns, and outliers [4].

Enter Big Data. As complex software systems have evolved into Software as a Service (SaaS) paradigm leveraging the advantages of economies of scale to make more powerful solutions, user experience analysis models have changed. Rather than having hundreds or thousands of users on a locally installed application, we now have millions of users working on a single common application. This single common application can record and track user activity and store the data in large-scale data warehouses creating an opportunity for a superior approach to user analysis.

Big Data analysis can be valuable beyond design and development; it can also play an important role in the way these features actually work. Leveraging real-time, the Related Search Suggestions feature encourages users to expand their queries which can lead to better research outcomes. And being data-driven, these features are rapidly and continuously fine-tuned to improve over time.

Relational database management system challenges and role of Big data

Hadoop and MapReduce seem to be more geared to situations where we are forced to large distributed scans of data, especially when those data aren't necessarily as homogeneous or as structured as what we find in the RDBMS world. The problem with an RDBMS is that in order to do this, we have to be really careful about how we structure our schema and partitions in order for it to work. Big Data architectures win when our data aren't structured enough to be partitioned and optimized easily or effectively in an RDBMS.

The limitations of traditional database architectures generally, they scale up with more expensive hardware, but have difficulty scaling out with more commodity hardware in parallel, and are limited by legacy software architecture that was designed for an older era. He contends that the Big Data era requires multiple new database architectures that take advantage of modern infrastructure and optimize for a particular workload. Examples of this are the C-store project, which led to the commercial database Vertical Systems, and the H-store project that led to VoltDB, an in-memory OLTP SQL database designed for high velocity Big Data workloads.

Big data applications: Real-world strategies for managing big data

Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation

Big Data and NoSQL: The Problem with Relational Databases

The NoSQL movement, where "NoSQL" stands for "Not Only SQL" is based on the concept that relational databases are not the right database solution for all problems. Relational databases are so ubiquitous in most organization these days that many people may not even be aware that there are other types of databases, let alone when using another database might be preferable. Relational databases perform transaction update functions very well, particularly handling the difficult issues of consistency during update. Production strength relational databases can handle the complexity of two phase commit capability, where one business transaction affects multiple databases and tables, and all updates have to be effected at the same moment [7].

However, relational databases apply much of the same overhead required for complex update operations to every activity, and that can handicap them for other functions. Relational databases struggle with the efficiency of certain operations key to Big Data management. Firstly, they don't scale well to very large sizes, and although grid solutions can help with this problem, the creation of new clusters on the grid is not dynamic and large data solutions become very expensive using relational databases. Secondly, they don't do unstructured data search very well (i.e. Google type searching) nor do they handle data in unexpected formats well. Thirdly, but not lastly, it is difficult to implement certain kinds of basic queries using SQL and relational databases, such as the shortest path between two points.

Social networking and Big Data organizations such as Face book, Yahoo, Google, and Amazon were among the first to decide that relational databases were not good solutions for the volumes and types of data that they were dealing with, hence the development of the Hadoop file system, the MapReduce programming language, and associated databases such as Cassandra and HBase. One of the key capabilities of a Hadoop type environment is the ability to dynamically, or at least easily, expand the number of servers being used for data storage. The cost of storing large amounts of data in a relational database gets very expensive, where cost grows geometrically with the amount of data to be stored, reaching a limit in the petabyte range. The cost of storing data in a Hadoop solution grows linearly with the volume of data and there is no ultimate limit.

I was a working programmer before relational databases were in common use. Yes, we did have electricity back then. And the databases I used were of the type called “hierarchical”. In fact, they were more efficient, in general, for high volume individual transaction processing than relational databases, although like relational databases they were not good for data that was structured inconsistently. But what we considered “high volume” then could be handled reasonably by my laptop now and those databases couldn’t handle dynamically allocating unlimited additional space, either [7].

Then comes the breed of NoSQL databases. I would treat them a subset of traditional RDBMS databases. Not all applications in this world will need all the functionality offered by RDBMS. If I want to use database as a cache, I would not care about durability. May be in some cases I would also not care about consistency. If all my data lookup is based on a key, I don’t need support for range queries. I may not need secondary indexes. I don’t need the whole query processing/query optimization layer which all the traditional databases have.

Hadoop And Mapreduce

According to IBM, 80% of data captured today is unstructured, from sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals, to name a few. All of this unstructured data is Big Data. Hadoop solve:

- Organizations are discovering that important predictions can be made by sorting through and analyzing Big Data.
- However, since 80% of this data is "unstructured", it must be formatted (or structured) in a way that that makes it suitable for data mining and subsequent analysis.
- Hadoop is the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes purposes.

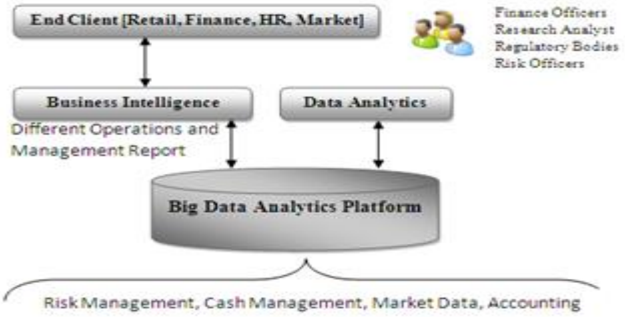

Fig 1. Big Data Analytics data flow architecture

Organizations are now creating more data than ever before, and as such a new set of tools and technologies are becoming popular to facilitate the storage & retrieval of this information in a timely and cost-effective manner. There are many technologies that are attempting to address these challenges, and as such there are different (and often incompatible) approaches, each with positives and negatives depending on the use-case.

While initially “Big Data” was synonymous with Hadoop, through aggressive vendor marketing and thought leadership discussion the term has broadened to mean “lots of data” and a wider set of data storage technologies. At a high-level, there are four competing sets of data storage/access technologies that you are likely to hear about related to big-data:

Explosive Market Dynamics

Market dynamics are changing due to big data. Data, like water, is powerful. Massive volumes of structured and unstructured data, wide variety of internal and external data, and high-velocity data can either power organizational change and business innovation, or it can swamp the unprepared. Organization that don’t adapt to big data risk [6]:

- Profit and margin declines
- Market share losses
- Competitors innovating faster
- Missed business opportunities

On the other hand, organizations that aggressively integrate big data thinking and capabilities will be able to:

- Mine social and mobile data to uncover customers’ interests, passions, associations, and affiliations
- Exploit machine data for predictive maintenance and operational optimization
- Leverage behavioral insights to create more a compelling user experience
- Integrate new big data innovations to modernize data warehouse and business intelligence environments (real-time, predictive)
- Become a data-driven culture
- Nurture and invest in data assets
- Cultivate analytic models and insights as intellectual property

Business And IT Challenges

ig Data enables business transformation, moving from a “rearview mirror” view of the business using a subset of the data in batch to monitor business performance, to the predictive enterprise that leverages all available data in real-time to optimize business performance. However, organizations face significant challenges in leveraging big data to transform their businesses, including [6]:

- Rigid architectures that impede exploiting immediate business opportunities
- Retrospective reporting that doesn’t guide business decisions
- Social, mobile, or machine insights that are not available in an actionable manner

Traditional business intelligence and data warehouses struggle to manage and analyze new data sources. Their architectures are:

- Batch-oriented which delays access to the data for analysis
- Brittle and labor intensive to add new data sources, reports, and analytics
- Performance and scalability challenged as data scales to petabytes.
- Limited to aggregated and sampled data views
- Unable to handle the tsunami of new, external unstructured data sources

Big Data Business Transformation

Where are an organization's aspirations with respect to leveraging big data analytics to power value creation processes? Some organizations struggle understanding the business potential of big data. They are unclear as to the different stages of business maturity. Our Big Data Maturity model benchmarks an organization's big data business aspirations, and provides a way to identify the level of sophistication desired for data monetization opportunities [6]:

- Business Monitoring – deploys business intelligence to monitor on-going business performance
- Business Insights – leverages predictive analytics to uncover actionable insights that can be integrated into existing reports and dashboards
- Business Optimization – embeds predictive analytics into existing business processes to optimize select business operations
- Data Monetization – creates new revenue opportunities by reselling data and analytics, creating "intelligent" products, or over-hauling the customer engagement experience
- Business Metamorphosis – leverages customers' usage patterns, product performance behaviors, and market trends to create entirely new business models

Big Data Journey

he big data journey requires collaboration between business and IT stakeholders along a path to identify the right business opportunities and necessary big data architectures. The big data journey needs to 1) focus on powering an organization's key business initiative while 2) ensuring that the big data business opportunities can be implemented by IT. The big data journey following this path [6]:

- Identify the targeted business initiative where big data can provide competitive advantage or business differentiation
- Determine – and envision – how big data can deliver the required analytic insights
- Define over-arching data strategy (acquisition, transformation, enrichment)
- Build analytic models and insights
- Implement big data infrastructure, technologies, and architectures
- Integrate analytic insights into applications and business processes

Operationalize Big Data

Successful organizations define a process to continuously uncover and publish new insights about the business. Organizations need a well-defined process to tease out and integrate analytic insights back into the operational system

The process should clearly define roles and responsibilities between business users, the BI/DW team, and data scientists to operationalize big data [6]:

- Collaborate with the business stakeholders to capture new business requirements
- Acquire, prepare, and enrich the data; acquire new structured and unstructured sources of data from internal and external sources
- Continuously update and refine analytic models; embrace an experimentation approach to ensure on-going model relevance
- Publish analytic insights back into applications and operational and management systems
- Measure decision and business effectiveness in order to continuously fine-tune analytic models, business processes, and applications

Value Creation City

Big data holds the potential to transform or rewire your value creation processes to create competitive differentiation. Organizations need a big data strategy that links their aspirations to the organization's key business initiatives. Envisioning workshops and analytic labs identify where and how big data can power the organization's value creation processes. There is almost no part of the organization that can't improve its value creation capabilities with big data, including [6]:

- Procurement to identify which suppliers are most cost-effective in delivering high-quality products on-time
- Product Development to identify product usage insights to speed product development and improve new product launches
- Manufacturing to flag machinery and process variances that might be indicators of quality problems
- Distribution to quantify optimal inventory levels and supply chain activities
- Marketing to identify which marketing campaigns are the most effective in driving engagement and sales
- Operations to optimize prices for "perishable" goods such as groceries, airline seats, and fashion merchandise
- Sales to optimize account targeting, resource allocation, and revenue forecasting
- Human Resources to identify the characteristics and behaviors of the most successful and effective employees

Exigency of cloud computing

While big data can yield extremely useful information, it also presents new challenges with respect to how much data to store, how much this will cost, whether the data will be secure, and how long it must be maintained [5].

The convergence of two key technological areas cloud computing and big data are having far reaching implications that indeed are changing the world.

Conclusion

The electronic age has provided enormous opportunities to advance our understanding of the world around us. Now we need to develop the storyboard to demonstrate the specific benefits and applications of Big Data Analytics for clients in capital markets, hedge funds and asset and wealth management, I immediately borrowed this line. We had already spent enough time with theoretical discussions, idea generation and brainstorming workshops. It was time to put the pedal to the metal and demonstrate real results. All complex problems and programs involve a learning curve and tackling Big Data is no exception. Some financial firms are re-thinking existing models and risk management analytics fuelled by readily available, open source Big Data technologies. Many cost effective, highly scalable, high performance and low latency Big Data Analytics tools became available in the last few years to assist in the collecting and loading of data from all data sources; from existing data warehouses to internal and/or external feeds as well as 3rd party data files. With the next generation analytics platforms investment management practitioners don't have to struggle for hours or days to create rich and realistic scenarios to analyze the impact of a certain market, security, or sector exposure on their investments as an event unfolds. They can quickly turn to a single place for instant, accurate information about their portfolio and track multiple dimensions of exposure data for the best course of action.

References

[1] IDC's Digital University Study, sponsored by EMC, Iune 2011

[2] Big - Data - Mining The differences, gains and application areas Peter Cochrane cochrane.org.uk ca-global.org 20 13

[3] Oracle : Big data for enterprise An Oracle White Paper June 2013 www.oracle.com/us/.../database/big-data-for-enterprise-519135.pdf

[4] Research challenges for data mining in science and Engineering http://www.cs.uiuc.edu/~hanj/pdf/ngdm09_han_gao.pdf

[5] Big Data: New Opportunities and New Challenges [Guest editors' introduction IEEE Journals & Magazines Publication Year 2013 Page(s): 22- 24

[6] The Big Data Story map https://infocus.emc.com

[7] http:/ /www.infocus.emc.com/

Author

Abhishek Khare received the MCA degrees in Computer Application from Shibaura Rajiv Gandhi Proudyogiki Vishwavidyalaya in 2007. Author has also published papers in Managing Risk in E-Commerce Security, Information Security Risk Analysis, Risk in E-Commerce, Emerging Standards of Data Mining in various National and International Journals.