Big Data: An Analysis of large data

Girish S. Thakare[1],Shilpa R. Deshmukh[2]
Assistant Professor: IT Dept.Sipna [1]
Assistant Professor: CSE Dept.Sipna COET [2]
Amravati, India
girish_thakare16@rediffmail.com,sharayu59@rediffmail.com

Abstract— **Big data is the emerging concept for large volume of data and complex structure with the solution to problems of storing, analyzing and visualizing for further processes. The process of research into massive amounts of data to reveal hidden patterns and secret correlations named as big data analytics. Big data handled the existing database management tool and give implementation to analyzed and execute the problem of massive data. This paper presents an overview of big data contents, scope, architectural structure and pattern**.

Keywords-Big data; database

I. INTRODUCTION

A decade ago, data storage scalability was one of the major technical issues data owners were facing. Nevertheless, a new brand of efficient and scalable technology has been incorporated and data management and storage is no longer the problem it used to be. In addition, data is constantly being generated, not only by use of internet, but also by companies generating big amounts of information coming from sensors, computers and automated processes. Big Data is a notion covering several aspects by one term, ranging from a technology base to a set of economic models.Big data is data that exceeds the processing capacity of conventional database systems [1]. The data is too big, moves too fast, or does not fit to the structures of database architectures. Big data can be stored, acquired, processed, and analyzed in many ways. Every big data source has different characteristics, including the frequency, volume, velocity, type, and veracity of the data. When big data is processed and stored, additional dimensions come into play, such as governance, security, and policies. Choosing an architecture and building an appropriate big data solution is challenging because so many factors have to be considered. The Big data architecture and patterns provides a structured and pattern-based approach to simplify the task of defining overall big data architecture. Big data problems are often complex to analyze and solve. The sheer volume, velocity, and variety of the data make it difficult to extract information and business insight. A good first step is to classify the big data problem according to the format of the data that must be processed [2], the type of analysis to be applied, the processing techniques at work, and the data sources for the data that the target system is required to acquire, load, process, analyze and store. To simplify the complexity of big data types, we can classify big data according to various parameters and provide a logical architecture for the layers and high-level components involved in any big data solution.

The new aspect of Big Data lies within the economic cost of storing and processing large datasets; the unit cost of storage has decreased by many orders of magnitude, amplified by the Cloud business model, significantly lowering the upfront IT investment costs for all businesses. As a consequence, the “Big Data concerns” have moved from big businesses and state research centers, to a mainstream status.

II. BIG DATA CORE CAPABILITIES

- Hadoop-based analytics: Processes and analyzes any data type across commodity server clusters.
- Stream Computing: Drives continuous analysis of massive volumes of streaming data with sub-millisecond response times.
- Data Warehousing: Delivers deep operational insight with advanced in-database analytics.
- Information Integration and Governance: Allows you to understand, cleanse, transform, govern and deliver trusted information to your critical business initiatives.

A. Supporting Platform Services

- Visualization & Discovery: Helps end users explore large, complex data sets.
- Application Development: Streamlines the process of developing big data applications.
- Systems Management: Monitors and manages big data systems for secure and optimized performance.
- Accelerators: Speeds time to value with analytical and industry-specific modules.

III. BIG DATA ARCHITECTURE AND PATTERNS

The big data provides several architectures and patterns to addresses the most common and recurring big data problems. The atomic patterns describe the typical approaches for consuming, processing, accessing, and storing big data. Composite patterns [3], which are comprised of atomic patterns, are classified according to the scope of the big data solution. Because each composite pattern has several dimensions, there are many variations of each pattern. The patterns enable business and technical users to apply a structured approach to establishing the scope and defining the high level solution for a big data problem.

A. Logical layers of a big data solution

Logical layers offer a way to organize your components. The layers simply provide an approach to organizing components that perform specific functions. The layers are merely logical; they do not imply that the functions that support each layer are run on separate machines or separate processes. A big data solution typically comprises these logical layers:

Big data sources: Think in terms of all of the data available for analysis, coming in from all channels. Ask the data scientists in your organization to clarify what data is required to perform the kind of analyses you need. The data will vary in format and origin:

- Format- Structured, semi-structured, or unstructured.
- Velocity and volume-The speed that data arrives and the rate at which it's delivered varies according to data source.
- Collection point-Where the data is collected, directly or through data providers, in real time or in batch mode. The data can come from a primary source, such as weather conditions, or it can come from a secondary source, such as a media-sponsored weather channel.
- Location of data source— Data sources can be inside the enterprise or external. Identify the data to which you have limited-access, since access to data affects the scope of data available for analysis.

Data massaging and store layer: This layer is responsible for acquiring data from the data sources and, if necessary, converting it to a format that suits how the data is to be analyzed. For example, an image might need to be converted so it can be stored in an Hadoop Distributed File System (HDFS) store or a Relational Database Management System (RDBMS) warehouse for further processing. Compliance regulations and governance policies dictate the appropriate storage for different types of data.

Analysis layer: The analysis layer reads the data digested by the data massaging and store layer. In some cases, the analysis layer accesses the data directly from the data source. Designing the analysis layer requires careful forethought and planning. Decisions must be made with regard to how to manage the tasks to Produce the desired analytics, Derive insight from the data, Find the entities required, Locate the data sources that can provide data for these entities, Understand what algorithms and tools are required to perform the analytics.

Consumption layer: This layer consumes the output provided by the analysis layer. The consumers can be visualization applications, human beings, business processes, or services. It can be challenging to visualize the outcome of the analysis layer. Sometimes it's helpful to look at what competitors in similar markets are doing.

B. Big data's three Vs

Input data to big data systems could be chatter from social networks, web server logs, traffic flow sensors, satellite imagery, broadcast audio streams, banking transactions, MP3s of rock music, the content of web pages, scans of government documents, GPS trails, telemetry from automobiles, financial market data, the list goes on. To clarify matters, the three Vs of *volume, velocity and variety* are commonly used to characterize different aspects of big data [4]. They're a helpful lens through which to view and understand the nature of the data and the software platforms available to exploit them.

1) Volume

Assuming that the volumes of data are larger than those conventional relational database infrastructures can cope with, processing options break down broadly into a choice between massively parallel processing architectures — data warehouses or databases. The benefit gained from the ability to process large amounts of information is the main attraction of big data analytics. Having more data beats out having better models: simple bits of math can be unreasonably effective given large amounts of data. If you could run that forecast taking into account 300 factors rather than 6, could you predict demand better? This volume presents the most immediate challenge to conventional IT structures. It calls for scalable storage, and a distributed approach to querying. Many companies already have large amounts of archived data, perhaps in the form of logs, but not the capacity to process it.

2) Velocity

It's not just the velocity of the incoming data that's the issue: it's possible to stream fast-moving data into bulk storage for later batch processing. There are two main reasons to consider streaming processing. The first is when the input data are too fast to store in their entirety: in order to keep storage requirements practical some level of analysis must occur as the data streams in. The second reason to consider streaming is where the application mandates immediate response to the data. Thanks to the rise of mobile applications and online gaming this is an increasingly common situation. The Internet and mobile era means that the way we deliver and consume products and services is increasingly instrumented, generating a data flow back to the provider. Online retailers are able to compile large histories of customers' every click and interaction: not just the final sales. Those who are able to quickly utilize that information, by recommending additional purchases, for instance, gain competitive advantage. The smartphone era increases again the rate of data inflow, as consumers carry with them a streaming source of geolocated imagery and audio data.

3) Variety

Rarely does data present itself in a form perfectly ordered and ready for processing. A common theme in big data systems is that the source data is diverse, and doesn't fall into neat relational structures. It could be text from social networks, image data, a raw feed directly from a sensor source. None of these things come ready for integration into an application. Even on the web, where computer-to-computer communication ought to bring some guarantees, the reality of data is messy. Different browsers send different data, users withhold information, they may be using differing software versions or vendors to communicate with you. And you can bet that if part of the process involves a human, there will be error and inconsistency. A common use of big data processing is to take unstructured data and extract ordered meaning, for consumption either by humans or as a structured input to an application. The process of moving from source data to processed application data involves the loss of information. When you tidy up, you end up throwing stuff away. This underlines a principle of big data: *when you can, keep everything*. There may well be useful signals in the bits you throw away. If you lose the source data, there's no going back.

IV. INTEGRATING BIG DATA WITH TRADITIONAL DATA

While the worlds of big data and the traditional data warehouse will intersect, they are unlikely to merge anytime soon. Think of a data warehouse as a system of record for business intelligence, much like a customer relationship management (CRM) or accounting system. These systems are highly structured and optimized for specific purposes. In addition, these systems of record tend to be highly centralized.

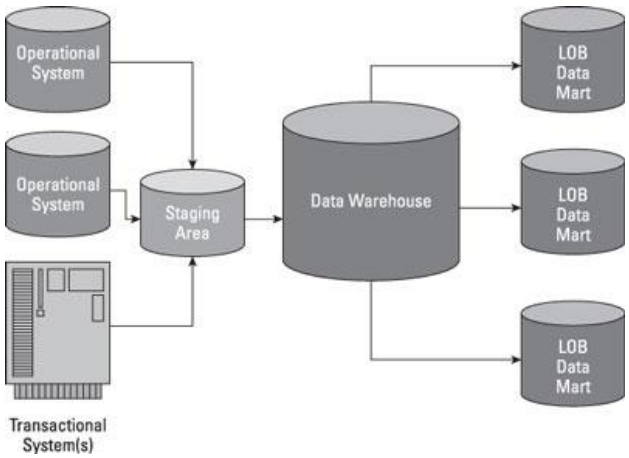

Figure 1. A typical approach to data flows with warehouses and marts.

Organizations will inevitably continue to use data warehouses to manage the type of structured and operational data that characterizes systems of record [5]. These data warehouses will still provide business analysts with the ability to analyze key data, trends, and so on. However, the advent of big data is both challenging the role of the data warehouse and providing a complementary approach. Think of the relationship

between the data warehouse and big data as merging to become a hybrid structure. In this hybrid model, the highly structured optimized operational data remains in the tightly controlled data warehouse, while the data that is highly distributed and subject to change in real time is controlled by a Hadoop-based(or similar NoSQL)infrastructure. It's inevitable that operational and structured data will have to interact in the world of big data, where the information sources have not (necessarily) been cleansed or profiled [6]. Increasingly, organizations understand that they have a business requirement to be able to combine traditional data warehouses with their historical business data sources with less structured and vetted big data sources. A hybrid approach supporting traditional and big data sources can help to accomplish these business goals.

V. SUMMARY

The data acquisition is increasing day by day using various sources, quality data is the big question today, Also managing such large volume of randomly generated data is big problem. For developers, layers offer a way to categorize the functions that must be performed by a big data solution, and suggest an organization for the code that must address these functions. For business users wanting to derive insight from big data, however, it's often helpful to think in terms of big data requirements and scope. Atomic patterns, which address the mechanisms for accessing, processing, storing, and consuming big data, give business users a way to address requirements and scope.

REFERENCES

[1] Marko Grobelnik ,Jozef Stefan Institute Ljubljana, Slovenia “Big-Data Tutorial” Stavanger, May 8th 2012.

[2] Thomas H. Davenport ,Jill Dyché “ Big Data in Big Companies” May 2013.

[3] Computational Social Science. David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer,Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. *Science* 6 February 2009: 323 (5915), 721-723.

[4] Big data: The next frontier for innovation, competition, and productivity. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. McKinsey Global Institute. May 2011.

[5] Materials Genome Initiative for Global Competitiveness. National Science and Technology Council. June 2011.

[6] Using Data for Systemic Financial Risk Management. Mark Flood, H V Jagadish, Albert Kyle, Frank Olken, and Louiqa Raschid. Proc. Fifth Biennial Conf. Innovative Data Systems Research, Jan. 2011.

[7] New Vantage Partners, “Big Data Executive Survey: Themes and Trends,” 2012.

[8] Richard L. Villars, Carl W. Olofson, Matthew Eastwood Big Data: What It Is and Why You Should Care June 2011.