www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 4 Issue 9 Sep 2015, Page No. 14378-14381

A Transformation from Relational Databases to Big Data

Shraddha Agarwal[1], Maddipatla Krishna Priyusha[2]

[1]JNTU-Hyderabad, Department of Computer Science and Engineering,
India,
shraddhaagarwal1994@gmail.com

[2]JNTU-Hyderabad, Department of Computer Science and Engineering,
India,
mk.priyusha@gmail.com

Abstract: ***We are living in an information age and there is enormous amount of data that is flowing between systems, internet, telephones, and other media. The data is being collected and stored at unprecedented rates. There is a great challenge not only to store and manage the large volume of data, but also to analyze and extract meaningful information from it. There are several approaches to collecting, storing, processing, and analyzing big data. The main focus of the paper is to draw an analogy for data management between the traditional relational database systems and the Big Data.***

Keywords: Big Data, Hadoop, NoSQL, Relational Database

1. Introduction

Data creation is occurring at an unprecedented rate. In 2010, the world generated over 1ZB of data; and by 2014, we have generated 7ZB of data. IBM estimates that every day 2.5 quintillion bytes of data are created – so much that 90% of the data in the world today has been created in the last two years. Increasingly large numbers of embedded sensors, smart phones, PCs, and tablet computers connected to network are generating enormous amounts of data. This data creates new opportunities to "extract more value" for the areas that it is needed. We have entered the age of "Big Data." Just as this data is generated by people in real time, it can be analyzed in real time by high performance computing networks, thus creating a potential for improved decision-making. The International Data Corporation (IDC) believes organizations that are best able to make real-time business decisions using Big Data solutions will thrive, while those that are unable to embrace and make use of this shift will increasingly find themselves at a competitive disadvantage in the market and face potential failure.[1]

2. Relational Databases

The traditional method of managing structured data includes a relational database and schema to manage the storage and retrieval of the dataset. For managing large datasets in a structured fashion, the primary approaches are data warehouses and data marts.

A data warehouse is a relational database system used for storing, analyzing, and reporting functions. The data mart is the layer used to access the data warehouse. A data warehouse focuses on data storage. The main source of the data is cleaned, transformed, catalogued, and made available for data mining and online analytical functions. The data warehouse and marts are Relational databases systems. The two main approaches to storing data in a data warehouse are the following:

- **Dimensional -** In a dimensional approach, transaction data are partitioned into "facts", which are generally numeric transaction data, and "dimensions", which are the reference information that gives context to the facts.
- **Normalized -** The tables are grouped together by subject areas that reflect data categories, such as data on products, customers, and so on. The normalized structure divides data into entities, which create several tables in a relational database.

Challenges in traditional data management using relational databases in an enterprise:

There are several challenges that the enterprises are faced today owing to the limitations posed by relational databases. Some of these are:

- Unstructured data that could provide a real-time business decision support remains unused as they cannot be stored, processed or analyzed.
- Several data islands are created and it becomes difficult to generate meaningful information from those.
- Data models are non-scalable and data becomes unmanageable.
- The cost of managing the data increases exponentially with the growth of data.

DOI: 10.18535/ijecs/v4i9.53

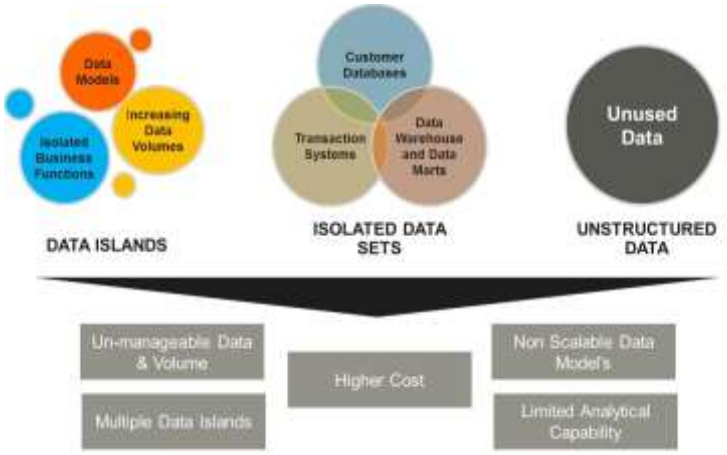

Figure 1: Traditional Data Management

3. Big Data

Big Data means Data sets whose volume and / or variety is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time that is relevant to business. The difficulty can be related to data capture, storage, search, sharing, analytics and visualization etc. It means data that's too big, too fast, or too hard for existing tools to process. Here, **"Too big"** means that organizations increasingly must deal with petabyte-scale collections of data that come from click streams, transaction histories, sensors, and elsewhere. **"Too fast"** means that not only is data big, but it must be processed quickly — for example, to perform fraud detection at a point of sale or determine which ad to show to a user on a webpage. **"Too hard"** means data that doesn't fit into an existing processing tool or that needs some kind of analysis that existing tools can't readily provide.

Big Data is characterized by the following 4 Vs:

- **Volume -** the vast amount of data generated every second that are larger than what the conventional relational database infrastructures can cope with.
- **Velocity** - the frequency at which new data is generated, captured, and shared.
- **Variety** - the increasingly different types of data (from financial data to social media feeds, from photos to sensor data, from video capture to voice recordings) that no longer fits into neat, easy to consume structures.
- **Veracity** - the disarrayed data (Facebook posts with hash tags, abbreviations, typos, and colloquial speech)

Big Data is all about

- Tapping into **diverse** data sets
- Discovering and co-relating **unknown** relationships within data
- Data driven insights for **faster** and **accurate** Business decisions

Big Data is a **capability**

- To **augment** enterprises' existing information fabric
- Solve today's **data problems**
- **Transform** the way business is done
- Build **competitive advantage** in the marketplace

4. Big Data Operational and Analytical Technologies

Though Big Data may sound futuristic, it does need certain exceptional technologies to efficiently process huge volumes of data in a good span of time. Here are some of the technologies that can be applied to the handling of big data.

Schema-less databases, or NoSQL databases - There are several approaches adopted by NoSQL (Not Only SQL) for storing and managing unstructured data. NoSQL databases separate data management and data storage, whereas relational databases combine both of them. One of the key concepts of the NoSQL is to have the database focus on the task of high-performance scalable data storage, and provide low-level access to a data management layer in a way that allows data management tasks to be written in the application layer rather than having data management logic spread across in SQL or DB-specific stored procedure languages

Most NoSQL databases can also be called schema-free databases. The key advantage of schema-free design is that it enables applications to quickly upgrade the structure of data without table rewrites. It also allows for greater flexibility in storing heterogeneously structured data. The data validity and integrity aspect is enforced at the data management layer.

NoSQL also has consideration in atomicity, consistency, isolation, and durability (ACID) aspects. It typically does not maintain complete consistency across distributed servers because of the burden this places on databases, particularly in distributed systems.

Massively Parallel Processing (MPP) - This involves a coordinated processing of a program by multiple processors (200 or more in number). Each of the processors makes use of its own operating system and memory and works on different parts of the program. Each part communicates via messaging interface. An MPP system is also known as
—loosely coupled or —shared nothing system.

Distributed file system or network file system allows client nodes to access files through a computer network. This way a number of users working on multiple machines will be able to share files and storage resources. The client nodes will not be able to access the block storage but can interact through a network protocol. This enables a restricted access to the file system depending on the access lists or capabilities on both servers and clients which is again dependent on the protocol.

Apache Hadoop is key technology used to handle big data, its analytics and stream computing. Apache Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers. It can be scaled up from a single server to thousands of machines and with a very high degree of fault tolerance. Instead of relying on high-end hardware, the resiliency of these clusters comes from the software's ability to detect and handle failures at the application layer.

Data Intensive Computing is a class of parallel computing application which uses a data parallel approach to process big data. This works based on the principle of collocation of data and programs or algorithms used to perform computation. Parallel and distributed system of inter-connected standalone computers that work together as a single integrated computing resource is used to process / analyze big data.

5. Big Data Framework

We have demonstrated a simple framework below to look at the key components of a Big Data system in order to work through many architectural decisions as you explore the world of big data. Big data often brings four new and very different considerations in enterprise architecture:

- **Data sources have a different scale** – many companies work in the multi-terabyte and some in petabyte arena.
- **Speed is critical** – nightly ETL (extract-transform-load) batches are insufficient and real-time streaming from solutions like s4 and Storm are required.
- **Storage models are changing** – solutions like HDFS (Hadoop Distributed File System) and unstructured data stores like Amazon S3 provide new options.
- **Multiple analytics paradigms and computing methods must be supported:**
    - **Real-time database and analytics:** These are typically in-memory, scale-out engines that provide low-latency, cross-data center access to data, and enable distributed processing and event-generation capabilities.
    - **Interactive analytics:** Includes distributed MPP (massively parallel processing) data warehouses with embedded analytics, which enable business users to do interactive querying and visualization of big data.
    - **Batch processing**: Hadoop as a distributed processing engine that can analyze very large amounts of data and apply algorithms that range from the simple (e.g. aggregation) to the complex (e.g. machine learning).

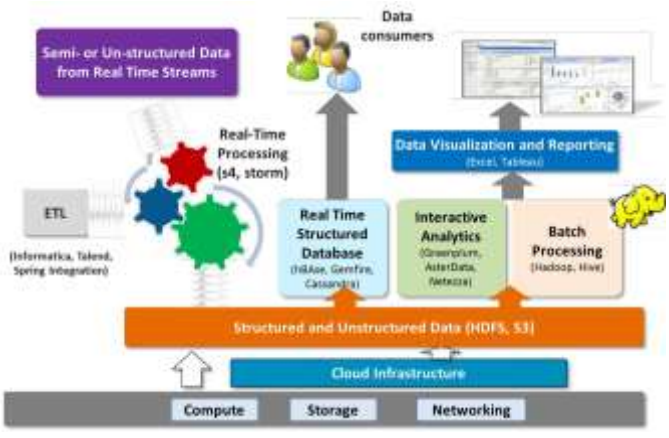

Figure 2: *Holistic View of a Big* Data Framework

6. Why transition from Relational databases to Big Data

The following table enunciates the difference between the traditional relational databases and Big Data database systems (Hadoop). Owing to the enormous amounts of data being generated and analyzed real time to provide intelligence to the decision support systems, there is a clear need of the time to transition to Big Data.

Table 1: Difference between RDBMS and Hadoop

	RDBMS	Hadoop
Description	Traditional row-column databases used for transactional systems, reporting, and archiving.	Distributed file system that stores large amount of file data on a cloud of machines, handles data redundancy etc. On top of that distributed file system, Hadoop provides an API for processing all that stored data - Map-Reduce. On top of this basic schema a Column Database, like hBase can be built.
Type of data	Works with structured data	Works with structured, semi-structured, and unstructured data
Max data size	Terabytes	Hundreds of Pitabytes
Limitations	Databases must slowly import data into a native representation before they can be queried, limiting their ability to handle streaming data.	Works well with streaming data
Read / write throughput limits	1000s queries/second	Millions of queries per second
Data layout	Row-oriented	Column family oriented

7. Suggested Big Data Adoption Roadmap for an Enterprise

Following is a suggested outline of the roadmap for adoption of Big Data for an Enterprise:

- **Architecture and Planning for Enterprise needs**
  - Define Business and Technology blueprints
  - Define entry points to Big Data
  - Create Governance framework
- **Build Foundation and Capabilities**
  - Create Development frameworks and methodology
  - Select Technology and Tools
  - Build Capability
  - Identify policy, privacy and security considerations
  - Build Proof of Concepts / Pilot
- **Adopt Big Data through Execution and Integration**
  - Operationalize Proof of Concepts / Pilot
  - Integrate Big Data into existing Information Management framework
  - Focus on Business value
  - Embed target state Enterprise capabilities in Business
- **Constant focus on Business value and Innovation through continuous improvement**
  - Continue to build and refine target state Enterprise capabilities
  - Continuous improvement and future readiness
  - Focus on Innovative Technology Solution

8. Conclusion

During the last 35 years, data management principles such as physical and logical independence, declarative querying and cost-based optimization have led, to a multi-billion dollar industry. More importantly, these technical advances have enabled the first round of business intelligence applications and laid the foundation for managing and analysing Big Data today. The many novel challenges and opportunities associated with Big Data necessitate rethinking many aspects of these data management platforms, while retaining other desirable aspects. We believe that appropriate investment in Big Data will lead to a new wave of fundamental technological advances that will be embodied in the next generations of Big Data management and analysis platforms, products, and systems. We believe that these research problems are not only timely, but also having the potential to create huge economic value for years to come. However, they are also hard, requiring us to rethink data analysis systems in fundamental ways. A major investment in Big Data, properly directed, can result not only in major scientific advances, but also lay the foundation for the next generation of advances in science,
medicine, and business.

According to IBM, 80 per cent of world‘s data is unstructured and most businesses don‘t even attempt to use this data to their advantage. Once the technologies to analyze big data reach their peak, it will become easier for companies to analyze massive datasets, identify patterns and then strategically plan their moves based on consumer requirements that identified through historic data. In conclusion then, big data will change the world with the ever-growing amounts of data and _large-scale analytics‘ (or simply _analytics‘ because what is large now will be normal tomorrow) in relation to our ability to analyze and harness big data.

References

[1] World's data will grow by 50X in next decade, IDC study predictshttp://www.computerworld.com/s/article/9217988/World_s_data_will_grow_by_50X_in_next_decade_IDC_study_predicts
[2] NoSQL Architecture a blog by Kris Zyp http://www.sitepen.com/blog/2010/05/11/nosql-architecture/
[3] From Databases to Big Data by Sam Madden – Article published in IEEE Internet Computing magazine
[4] Considerations for Big Data: Architecture and Approach by Kapil Bakshi – Paper published in IEEE
[5] Big Data Analytics by Sachidanand Singh – Paper published in 2012 International Conference on Communication Information & Computing Technology (ICCICT), Oct. 19-20, Mumbai, India
[6] McKinsey Global Institute
[7] http://hadoop.apache.org/