SRI AKILANDESWARI WOMEN’S COLLEGE, WANDIWASH

BIG DATA ANALYTICS

Class: II. M. Sc Computer Science

Prepared by
C. BALASUBRAMANIAN,
Assistant Professor, Dept of Computer Science

SWAMY ABEDHANADHA EDUCATIONAL TRUST, WANDIWASH

New Approach to Enterprise Information Management [EIM] for Big Data

Type of Data:

- relational, object, network Database
- In the big data scenario, the EIM needs to manage all kinds of data, including traditional structured data, semi-structured, unstructured and poly-structured data, and content such as e-mails, web-page content, video, audio, text, graphical representations,etc

Enterprise data modeling

- The cost of scaling and managing infrastructure while delivering a satisfactory consumer experience for newer applications such as web 2.0 and social media applications has proven to be quite steep. This has led to the development of “NoSQL” databases as an alternative technology with features and capabilities that deliver the needs of the particular use case.

Data Integration

To overcome the challenges in the big data scenario, there has been a push toward focusing on extract and load approaches

Cost

there are several new technologies and architectures enabling companies with cost effective solutions. how it solves the big-data-related issues while at the same time providing a cost effective viable alternative to IT infrastructure.

Data Quality

it is recommended that ongoing data quality initiatives be focused on resolving data quality issues for transactional and reference/master data either closer to the source and/or downstream. For the big data scenarios, there is tremendous value in applying data quality rules to the big data sets and getting an idea of the conformance of such data sets to the applied rules.

Master database management [MDM]

- The biggest advantage of big data sources is that they help in validating your master entities and in many cases help in enriching them. For example, using Google e-mail ID, Facebook IDs and LinkedIn IDs you can further enrich you customer identification process and improve your conversations with customers through multiple channels.

Metadata Management

- when you are dealing with big data sources, you may not find well-documented definitions associated with data attributes. This is precisely why you should attempt to create a minimum set of documentation consisting of the source, how you accessed it, what access methods (APIs or direct downloads) you applied, what data cleansing methods you applied, what security and privacy measures you applied on the data sets, where you are storing the raw data sets,

Skills

- In big data scenarios, data scientists and data architects rather than database administrators will be in demand to effectively implement the distributed nature of big data processing, ingesting and aggregating data from multiple sources and managing storage, compute, and network resources to handle large data sets

New capabilities needed for big data

- **Data Discovery:** consists of activities involving locating, cataloging, and setting up access mechanisms for data sources.
  Such an exercise greatly benefits the enterprise in agile data integration, enriching the content and value of enterprise data assets from both internal and external data sources.

- **Rapid Data Insight:** is the next generation of agile data analysis wherein data from multiple sources can be quickly inspected, cleansed, and transformed with the goal of getting a deeper understanding of the data, spot apparent trends and patterns, and getting an idea of the value of data assets in supporting decision making and analytics. Data insight enables end users to make better "sense" of data assets.

- **Advanced Data Visualization:** is the process whereby reliable data from one or more sources are integrated or mashed up together and visually communicated clearly and effectively through advanced graphical means. This enables you to succinctly present and convey the insight gleaned from large amounts of information and enables better cognitive understanding of such information insight especially to business end users.

- **Advanced Analytics:** involves the application of business rules, domain knowledge and statistical models, often in-database closer to the data sources themselves, that help in decision making and help answer the questions of “What?” and “Now What?”.

- **Data Virtualization:** is a data integration technique that provides complete, high-quality, and actionable information through virtual integration of data across multiple, disparate internal and external data sources. Instead of copying and moving existing source data into physical, integrated data stores (e.g., data warehouses and data marts), data virtualization creates a virtual or logical data store to deliver the data to business users and applications

- **Data Services:** is described as a modular, reusable, well-defined, business-relevant service that leverages established technology standards to enable the access, integration, and right-time delivery of enterprise data throughout the enterprise and across corporate firewalls. Data services technology provides an abstraction layer between data sources and data consumers