PGDBD-Core course III

IT 603 Data warehousing and integration

Introduction: Data warehousing: Definition, need and milestones of data warehousing.

Transactional systems and operational data storage.

Data warehouses and data marts. Architectures of a data warehouse. Generalized, federated and hub and spoke data warehousing. Architecture framework. Management and control modules.

Components of a data warehouse. Source data components. Data staging and storage components. Data delivery and meta data components. Management and control components.

Real time data warehousing. Query tools. Browsing tools. ERP, KM and CRM data warehousing. Active data warehousing.

Data fusion and integration. Data standards. Metadata and OLAP. Web-enabled data warehousing.

Workload management in data warehouse. Query classification, ETL and CDC workloads. Data loading techniques. Data transport.

Cloud computing. Infrastructure as a service. Platform as a service. Data warehousing on a cloud computer-technologies and issues.

Data virtualization. In-memory technologies.

Integration strategies in data warehouse. Data driven integration. Physical component integration and architecture. External data integration.

Hadoop and RDBMS. Semantic framework for data integration. Lexical processing. Semantic knowledge processing. Information Curation. Visualization.

Biological data integration using InterMine framework. InterMine powered data warehouses. FlyMine, MouseMine, YeastMine, INDIGOmine, TargetMine, HumanMine, PhytoMine for plant data.

Biological data warehousing with BioMart. ID conversion.
Biological Sequence retrieval and enrichment analysis.

Practicals:

1. Setting up a batch queuing system such as Torque from scratch and configuring its multiple functionalities.
2. Using Apache Spark and Hadoop to perform tasks on a pre-configured Cloud.