



CLOUD COMPUTING APPLICATIONS

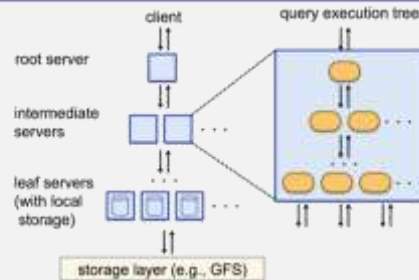
Analytics in the Cloud: Modern Data Warehouses
Prof. Reza Farivar

Modern Data Warehouse Architecture

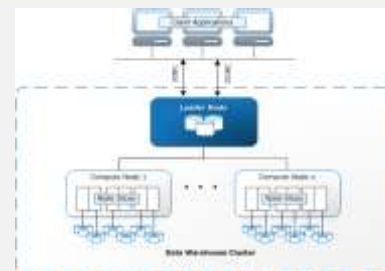
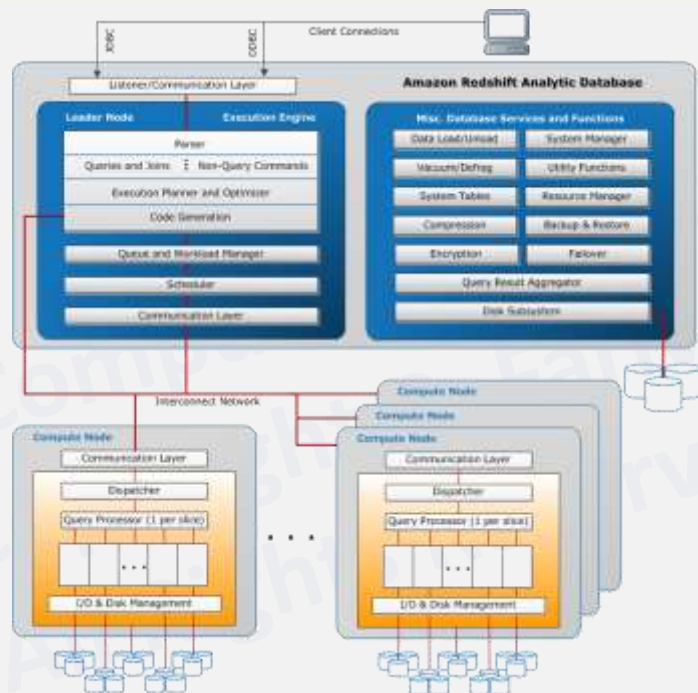
- **Cloud**
 - access to near- infinite, low-cost storage
 - improved scalability
 - Outsourcing of data warehousing management and security to the cloud vendor
 - Pay per use
- **Massively parallel processing (MPP)**
 - Dividing computing operations to execute simultaneously across many separate computer processors
 - Like Sharding
- **Columnar storage**
- **Vectorized processing**

Columnar-based Data Warehouses

- Column Store, MPP, Cloud based
- MariaDB with InfiniDB
 - For reference: row based regular engine for OLTP: InnoDB
- PostgreSQL
 - Citus cstore_fdw
- Google BigQuery
 - Based on Google Dremel, paper published in 2010
- AWS Redshift
 - Based on an older version of PostgreSQL
 - PostgreSQL 8.0.2
 - Originally developed by ParAccel
 - Some PostgreSQL features that are suited to smaller-scale OLTP processing, such as secondary indexes and efficient single-row data manipulation operations, have been omitted to improve performance



Redshift Architecture



Microsoft Azure Synapse

- Azure Synapse Analytics
 - Formerly Azure SQL Data Warehouse
 - SQL Analytics: Complete T-SQL based analytics
 - SQL pool (pay per DWU provisioned)
 - SQL on-demand (pay per TB processed)
 - Spark: Deeply integrated Apache Spark
 - Integration with Power BI

