CLOUD COMPUTING APPLICATIONS

# Analytics in the Cloud: Data Lake

Prof. Reza Farivar

# Data Lake

- Data Warehouses cannot accommodate unstructured big data projects
  - Petabytes of data in structured, semi-structured and unstructured forms
  - Semi-structured and unstructured data: JSON, XML, Log files, Natural Language, Images, video, etc.
    - Social media sites, mobile phones, Internet of Things (IoT) devices, and many other sources, including shared data sets
  - Structured data typically collected from enterprise applications
- Data Lake: a new type of data repository for storing massive amounts of raw data in its native form, in a single location
  - "A large body of water, into which new water streams from many channels, and from which samples are taken and analyzed"
  - Solution to a growing problem: the need for a scalable, low-cost data repository that allowed organizations to easily store ***all*** data types and analyze that data to make evidence-based business decisions
- The initial data lakes were deployed on premises, mostly using open source tools from the Apache Hadoop ecosystem
- Modern data lakes combine the power of analytics with the flexibility of big data models and the agility and limitless resources of the cloud

# Components of a Modern Data Lake

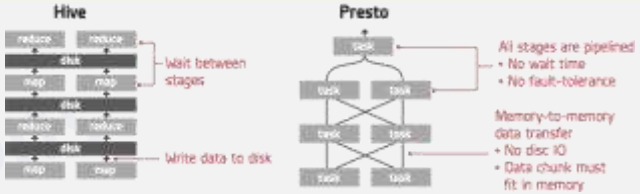
- Object Storage to store all data types
  - Big Data: Azure Data Lake Storage
- Move Data
  - AWS Data Pipeline
  - Azure Data Factory
- Data Lake Schema Discovery:
  - A fully managed service that serves as a system of registration and system of discovery for enterprise data sources
  - AWS Glue
  - Azure Data Catalog
- SQL Exploration and Query
  - Apache Presto
  - AWS Athena
- Lake Formation
  - AWS Lake Formation
  - Azure Data Share

# Discovery

- Data Lake Discovery Services
- Data Crawler
- Metadata extraction → Schema → Catalog
- ETL workloads
- Apache Atlas
- AWS Glue
  - Serverless
  - Data Sources: Amazon Redshift, Amazon S3, Amazon RDS, and Amazon DynamoDB
  - AWS Glue Data Catalog: Crawls your data sources, identifies data formats, and suggests schemas and transformations
  - AWS Glue ETL:
    - AWS Glue provides a number of built-in preload transformations that let ETL jobs modify data to match the target schema
    - Automatically generates code to execute data transformations and loading processes for more complex, custom ETL transformations
  - ETL jobs on a fully managed, scale-out Apache Spark environment
- AWS Data Pipeline: Focus on data transfer
- Azure Data Catalog: Discovery
- Azure Data Factory: ETL
- Google Cloud Data Catalog, Dataflow

# Data Lake Exploration and Query

- Directly run SQL queries on the data lake
- No need to setup intermediary databases or data warehouses
- Apache Presto
  - In-memory distributed SQL query engine
  - Optimized for star schema joins
    - 1 large Fact table and many smaller dimension tables
  - Interactive Hive on Steroids
- Aws Athena
  - Managed Serverless Presto
- Azure Data Lake Analytics
- Apache Spark SQL
  - IBM Cloud SQL Query




```
@a  =
    SELECT * FROM
        (VALUES
            ("Contoso", 1500.0f),
            ("Woodgrove", 2700.0f)
        ) AS
            D( customer, amount );
OUTPUT @a
    TO "/data.csv"
    USING Outputters.Csv();
```

*Example Azure Data Lake Analytics U-SQL query on a Data Lake output goes to a csv file on the data lake*

# Cloud-based Data Lake Automation

- Automated tools to orchestrate the data transfer, discovery, ETL and analytics steps
- AWS Lake Formation
  - Glue
  - Athena
  - Redshift Spectrum
  - EMR
    - Apache Zeppelin or EMR Notebooks
- Azure Data Share