Day 2: Big Data and Azure Introduction

Wednesday, August 30, 2023 9:31 AM

What is Big data?
Big data is high volume, high velocity, veracity (uncertainty of data) and high variety information assets that require new f orms of processing to enable enhances decision making, insight, discovery and process optimization.

Batch and stream data processing

Batch processing
Recorded events ----read----> periodic query/application ----write----> database/HDFS and a report

Stream processing
Real time events ----ingest----> continuous query/application ----update----> database/k-v store ----read----> live report/dashboard

How Stream analytics helps e commerce?
- Get 360 degree view of customer, what they require and need from the company
- Recommendations based on what the consumer buys
- Use sentiment analysis to go through reviews and get to know how the products are doing

Parallel and Distributed Processing

Parallel Processing
Parallel processing is a method in computing of running two or more processors (CPUs) to handle separate parts of an overall task. Breaking up different parts of a task among multiple processors will help reduce the amount of time to run a program. Any system that has more than one CPU can perform parallel processing, as well as multi-core processors which are commonly found on computers today.

Distributed Processing
Distributed processing means that a specific task can be broken up into functions, and the functions are dispersed across two or more interconnected processors. A distributed application is an application for which the component application programs are distributed between two or more interconnected processors. Distributed data is data that is dispersed across two or more interconnected systems.

Cloud Benefits
- High availability
- Scalability
- Global reach
- Agility
- Disaster recovery
- Fault recovery
- Elasticity
- Customer latency capabilities
- Predictive cost considerations
- Security

Types of clouds
- Private Cloud: No access to users outside the organization
- Public Cloud: Owned by Cloud services, available via secure network, upscale and downscale on demand
- Hybrid cloud: combines public and private cloud

Serverless Computing
Serverless is a cloud computing application development and execution model that enables developers to build and run application code without provisioning or managing servers or backend infrastructure.

Data Warehouse
A data warehouse is a central repository of information that can be analyzed to make more informed decisions. Data flows into a data warehouse from transactional systems, relational databases, and other sources, typically on a regular cadence.
- Save data as object
  - As table or view
- Structured data
- Dimension modeling
- Fixed schema
- Follows ACID properties
  - Atomicity
  - Consistency
  - Isolation
  - Durability

Data Lake
A data lake is a centralized repository designed to store, process, and secure large amounts of structured, semi-structured, and unstructured data. It can store data in its native format and process any variety of it, ignoring size limits.
- Flexible schema
- Does not follow ACID properties
- Blob Storage: Binary large object, can store all big data in binary

Lakehouse
- Combined advantages of Data Warehouse and Data Lake
- Follows ACID properties and can simultaneously store all kind of data types

CapEx
Capital expenditures (CapEx) are funds used by a company to acquire, upgrade, and maintain physical assets such as property, plants, buildings, technology, or equipment

OpEx
An operating expense is an expense that a business incurs through its normal business operations. Often abbreviated as OpEx, operating expenses include rent, equipment, inventory costs, marketing, payroll, insurance, step costs, and funds allocated for research and development

Resource Group
Container that holds all your resources in one place, example storage, web, VM etc.