Big Data

Wednesday, August 30, 2023 9:06 AM

- **What is Big Data?**
- Data that is huge in volume, variety, velocity, veracity (Uncertainty of data, data in motion, or at rest)
- Batch vs Stream Processing: streaming data is immediate, constant sending of data, real time events
  Batch takes time, recorded events
- **Stream Analytics:**
  - get 360 degree view of customer, what they require and need from the company
  - Recommender Systems: if pant is bought, t-shirt recommended
  - Sentiment Analysis: collect feedback from customer, if positive negative neutral
- **Hadoop:** framework for batch processing, distributed processing
- **Spark:** for stream processing, supports multiple languages, we will learn pyspark
  - Parallel vs Distributed Processing: in parallel, multiple processes are being completed side by side, in distributed, multiple computers work on the same process side by side.

Cloud Benefits: High availability, scalability, fault tolerance, elasticity (can expand), global reach, agility (responding to market change, new trends there in public cloud), predictive cost consideration, disaster recovery, security.

Capex: Expenditure of everything, infra network engineer salary, maintenance, etc.
Opex: pay as you go, get billed immediately, spend on products and services as needed. [consumption based model]

Data Warehousing:
- Can store data as object
- Structured Data
- Dimensional Modeling
- Fixed schema
- Acid

Data Lake:
- Can store all semi, unstruct, struct
- Flexible schema
- Acid property not available: atomicity, consistency, isolation, durability
- no need to fill all properties such as name, id, location, can just put name
- Blob storage: binary large object, stores all big data in binary

IaaS: only skeleton needed, eg. Server, machines etc. so only infrastructure needed.
PaaS: till runtime env, everything given but application and data and access is your own.
SaaS: Only data and access is your own, rest provided.

Lakehouse:
- Combined advantages of dwh and datalake
- Has acid properties as well as all kinds of data can be used

Resource group: container that holds all your resources in one place, for example, Storage, VM, web and db.

Private Cloud: No access to users outside organization.
Public Cloud: owned by cloud services, available via secure network, upscale and downscale on demand
Hybrid Cloud: combines public and private cloud. When to use?

Serverless Computing: cloud provider automatically provisions scales and manages infra required to run the code. Eg. Azure functions and Azure Apps.