Big Data

Wednesday, August 30, 2023
9:23 AM

BIG DATA

Data sets that are too large or complex to be dealt with by traditional data-processing application software.

Characteristics:

i. ***Volume:*** Size of data plays a very crucial role in determining value out of data.
ii. ***Variety:*** heterogeneous sources and the nature of data, both structured and unstructured.
iii. ***Velocity:*** speed of generation of data.
iv. ***Variability:*** inconsistency which can be shown by the data at times.

Batch Processing:

- Processing of high volume of data in batch within a specific time span.
- Used when data size is known and finite.
- Takes longer time to processes data.

Stream Processing:

- Processing of continuous stream of data immediately as it is produced.
- Used when the data size is unknown and infinite and continuous.
- Takes few seconds or milliseconds to process data.

PARALLEL COMPUTING

- Single computer uses multiple processors to process tasks in parallel.
- Shared or distributed memory.
- Processors communicate with each other through bus.
- Improves the system performance.

DISTRIBUTED COMPUTING

- Uses multiple computing devices to process tasks.
- Distributed memory.
- Computers communicate with each other through message passing.
- Improves system scalability, fault tolerance and resource sharing capabilities.

DATA WHAREHOUSE

Store data as Object

1. Structured Data
2. Dimension modelling.
3. Fixed schema
4. ACID properties.

DATA LAKE

- Whatever data you want to store you can store.
- Flexible schema
- No ACID properties

LAKE HOUSE

Having advantages of both data warehouse and data lake.

AZURE FUNDAMENTALS

- Capex: Capital expenditure, infrastructure.
- Opex : only pay for operations . Only pay when needed.
- ARM: management architecture.
- Resource: services that you get from azure.
- Resource groups: Logical container