BIG DATA

How data becomes Big Data

* Follows 5 V’s
* Volume
* Velocity
* Veracity
* Variety
* Value

Types of Big Data Processing

* Batch Processing
  + Data is stored for longer intervals before being processed
* Stream Processing
  + Data is stored in short-term storage before being processed

Working on Big data

* Extracting data and carrying out transformations

How stream analytics helps a business

* Customer requirements can be easily anticipated
* Better recommendations can be made
* Marketing a product becomes easier

Sentiment Analysis

* To evaluate customer reaction to a certain product
* Using feedback data smart decisions can be made to create greater value for the company

Distributed computing and Parallel Processing

* To work with large amounts of data, the data is spread across multiple clusters each having multiple nodes which process the data in parallel, thus, making the operation much more efficient

ACID properties

* Atomicity
  + Either the entire operation happens or none of it happens
* Consistency
  + The data must remain the same across all of its copies
* Isolation
  + Every transaction is independent from other transactions
* Durability
  + The data remains the same for as long as required until some changes are made by the user

Data Warehousing

* Saves data as objects
* Stores only structured data
* Allows dimension modelling
* Has a fixed schema
* Follows ACID properties

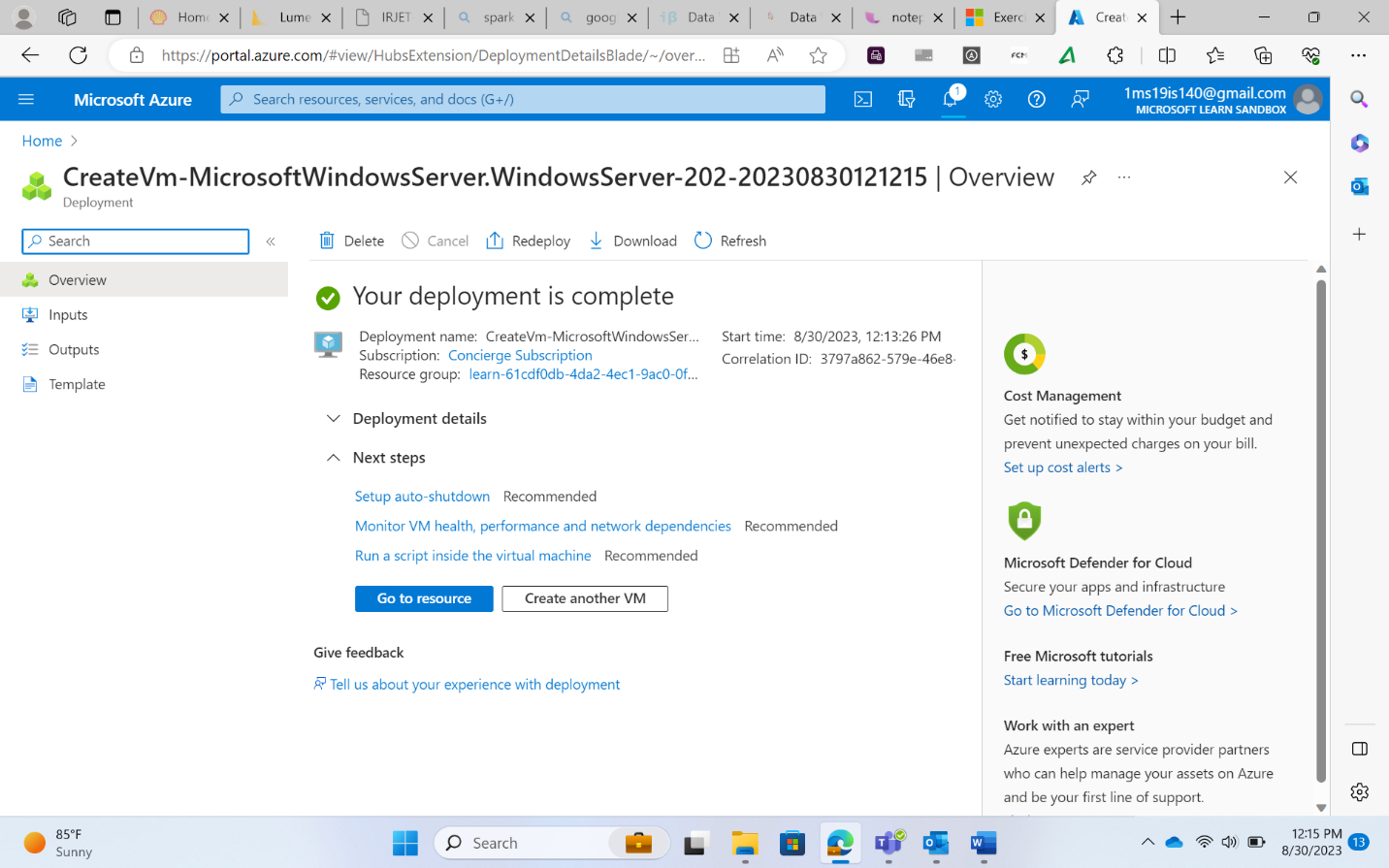
Data lake

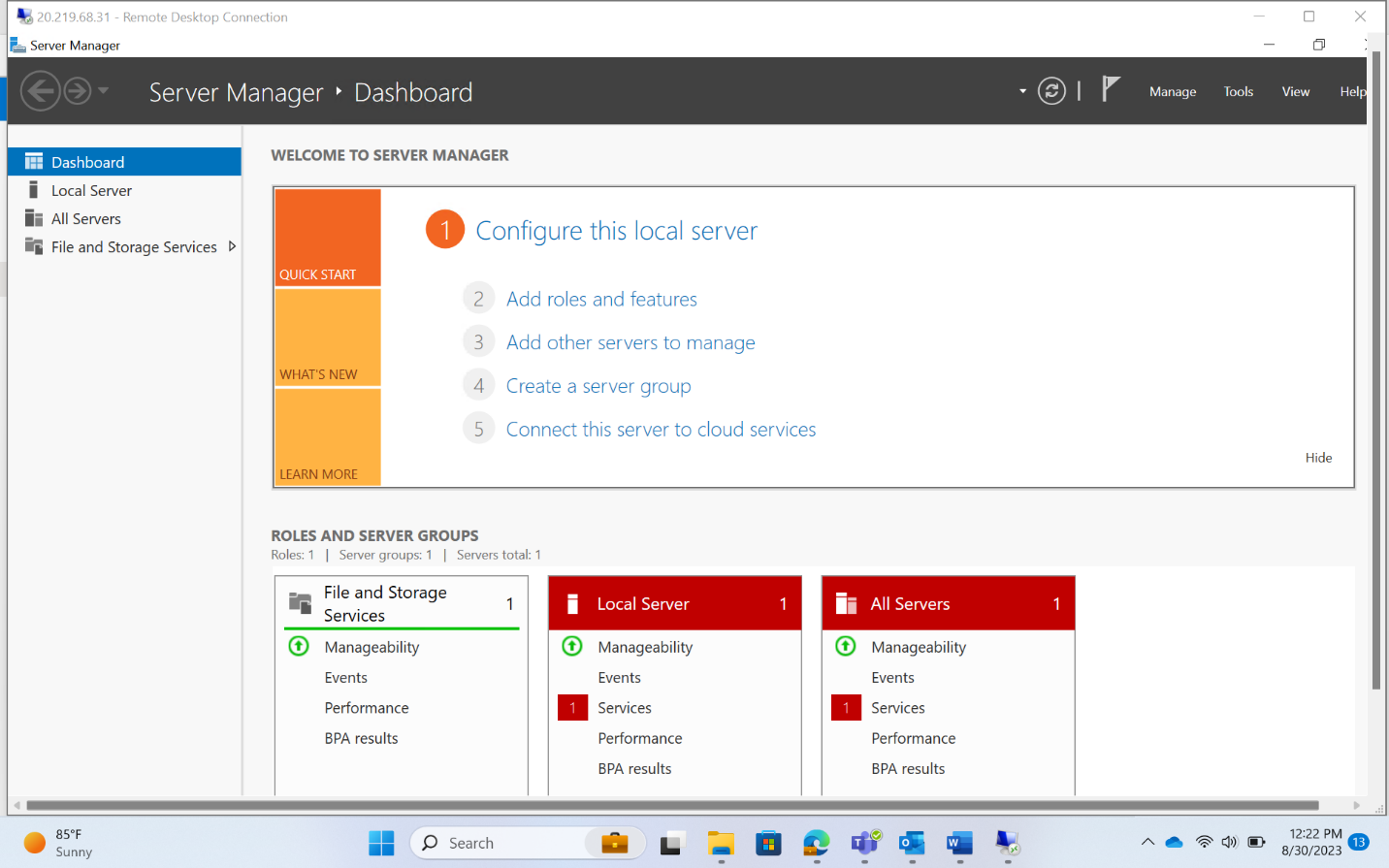
* Semi-structured and unstructured data can be stored here unlike in warehouses
* Does not follow ACID properties
* Uses BLOB objects to store data

Data Lakehouse

* Takes the best properties of warehouses and lakes
* Follows ACID properties
* Allows storage of all types of data

Azure VM Creation using Sandbox





Azure Certification

* DP-203
* It is an associate level certification
* Not needed to do fundamental certification as well

Cloud Benefits

* High availability
* Fault tolerance
* Scalability
* Elasticity
* Global reach
* Customer latency capabilities
* Agility
* Predictive cost consideration
* Disaster recovery

CapEx

* Capital expenditure
* Pay for the entire infrastructure

OpEx

* Operational Expenditure
* Pay only for the services used

IaaS

* Infrastructure as a service
* Compute, network, storage are provided by the cloud service provider

PaaS

* Platform as a Service
* Development and deployment of applications on the cloud

SaaS

* Software as a Service
* Outlook, One Drive, google drive, etc.

Serverless computing

* Azure Server
* Azure Logic Apps

AZURE CLOUD COMPUTING

Region pairs

* At least 300 miles between the data centers in a pair
* Automatic replication for some services
* Updates are made sequentially to the data centers in a pair

Availability Zones

* Zone 1, 2, 3
* To safeguard against downtime
* Connection between zones done using fiber optic network

Azure resources

* VM
* Storage
* Virtual accounts
* App services
* SQL DB
* Functions

Azure Subscription

* To avail Azure services, a subscription is needed
* Subscription: Development, Test, Production

Azure Resource Manager

* CRUD for Azure Services

\*\*\* Any service availed from Azure is called Resource

Resource Groups

* A logical container
* It holds all the resources

\*\*\* Subscription -> Resource Groups -> Resources

* Inside Analytics category, all the data engineer-oriented tools are present
* If a resource group already has a subscription, then all accounts connecting to that resource group can avail the subscription
* Tags are similar to social media tags. They are additional information about a resource
* 2 things to look out for
  + Resource group
  + Subscription