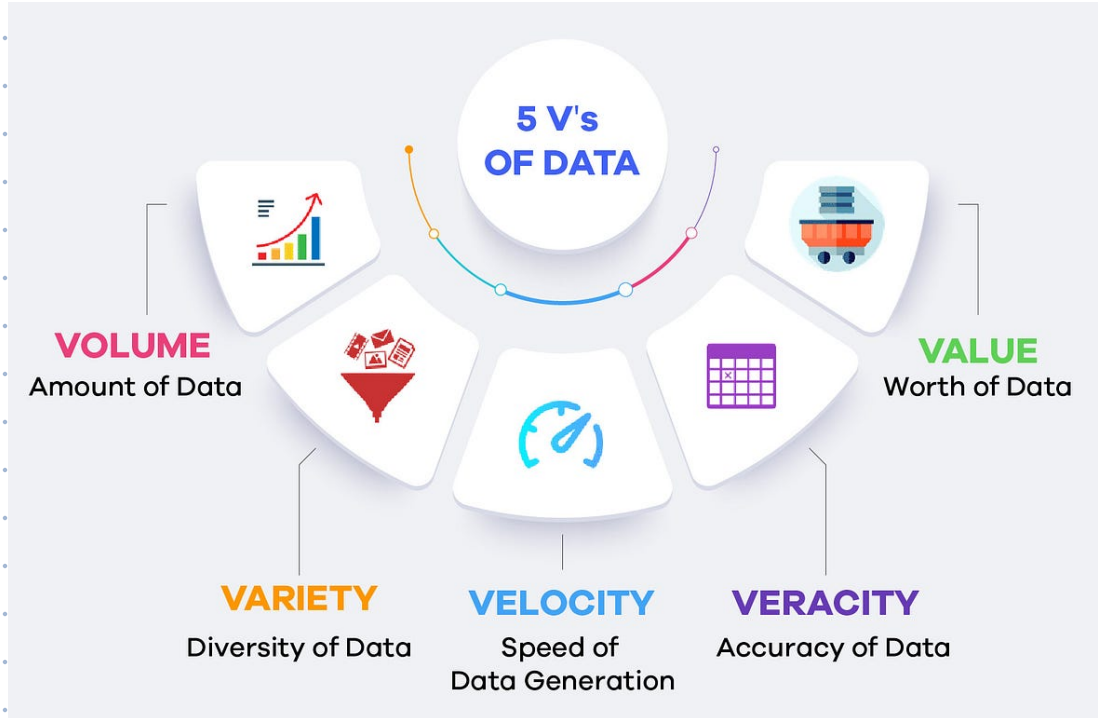
DATA – Raw facts, Images etc



1. Volume

The total amount of data generated, stored, and processed, ranging from gigabytes to petabytes and beyond.

Aspect	Definition
Small-scale Data	Manageable datasets that can be processed on a single machine (e.g., local databases, small CSV files).
Large-scale Data	Massive datasets requiring distributed computing (e.g., big data frameworks like Hadoop, Spark).

2. Variety

The different types and formats of data.

Type	Definition
Structured Data	Data organized in fixed formats, such as tables (e.g., SQL databases, spreadsheets).
Unstructured Data	Data without a predefined structure, such as images, videos, or social media posts.
Semi-structured Data	Data with some structure, often stored in formats like JSON or XML files.

3. Velocity

The speed at which data is generated, processed, and analyzed.

Type	Definition
Batch Processing	Data processed in large sets or chunks over a specific time frame (e.g., ETL jobs).
Real-time Data	Data processed immediately as it arrives (e.g., stock prices, live tracking).
Streaming Data	Continuous, real-time data flow from sensors or IoT devices.

4. Veracity

The accuracy, quality, and reliability of data, ensuring it is trustworthy for decision-making.

Type	Definition
Accurate Data	Reliable, precise, and error-free data that can be confidently used for analysis.
Biased Data	Data that may contain inaccuracies or distortions due to human or systemic biases.
Noisy Data	Data with irrelevant or misleading information that can obscure meaningful insights.

5. Value

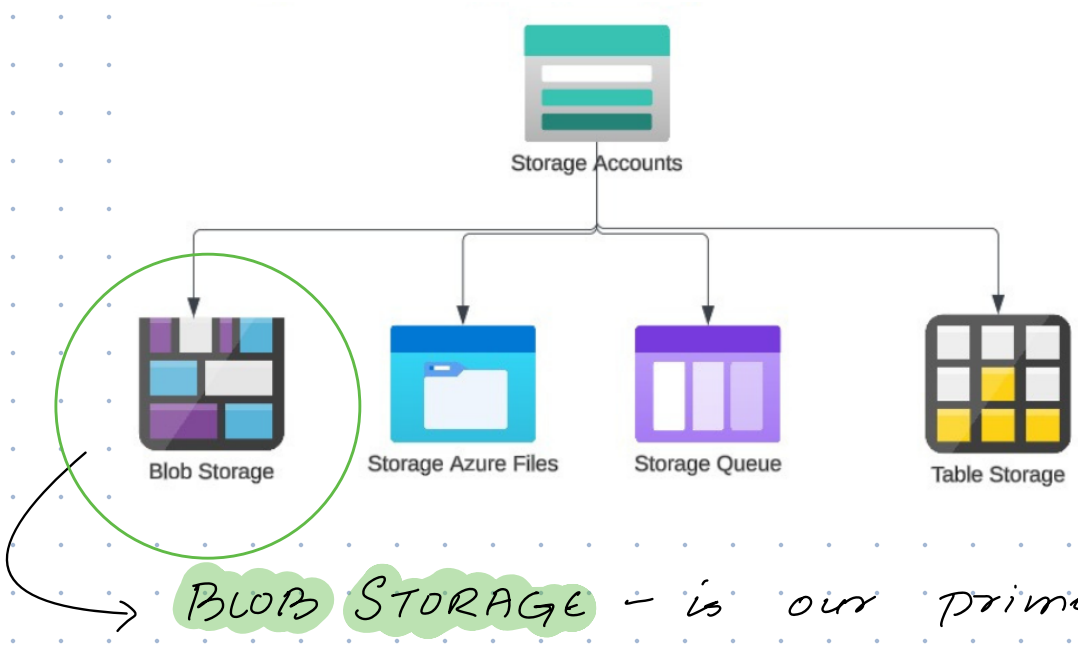
The usefulness of data in generating insights, making informed decisions, and driving business impact.

Type	Definition
Business Insights	Actionable insights derived from data, helping organizations optimize operations and strategy.
Predictive Models	Data-driven algorithms that forecast future trends or outcomes.
Optimization	Using data to enhance efficiency, improve processes, and maximize performance.

VARIETY OF DATA

Category	Data Type	Definition
Structured Data	**Relational Databases**	Data stored in **tables with rows and columns**, following a strict schema (e.g., SQL databases).
	Spreadsheets	Tabular data used for analysis and reporting (e.g., Excel, Google Sheets).
	ERP & CRM Systems	Business applications managing structured data for **operations and customer relationships** (e.g., SAP, Salesforce).
Semi-Structured Data	**JSON & XML Files**	Data with **key-value pairs or tags**, providing partial structure but no fixed schema (e.g., API responses, config files).
	NoSQL Databases	Schema-less databases storing data in **key-value, document, column, or graph models** (e.g., MongoDB, Cassandra).
	CSV Files	Text files storing **tabular data** without enforced schema.
	Parquet & ORC	**Optimized columnar storage formats** for big data, enabling **efficient compression and fast querying** (e.g., Apache Parquet, ORC).
Unstructured Data	**Images & Videos**	Media files lacking a predefined structure, requiring metadata for organization (e.g., PNG, MP4).
	Audio Files	Sound recordings used in **voice assistants, music streaming, and call logs** (e.g., MP3, WAV).
	Emails & Social Media	Text-heavy, unstructured data with mixed content like **attachments, messages, and comments** (e.g., Outlook, Twitter posts).

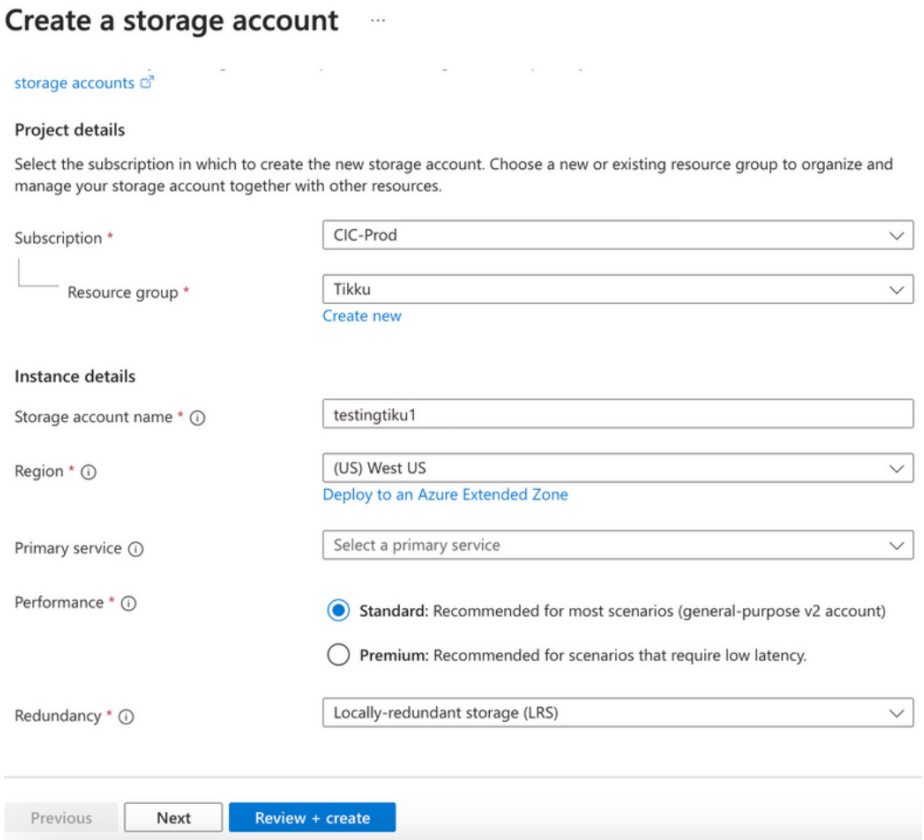
Azure Storage Services;



Blob Storage - is our prime focus

Storage Type	Description	Best Use Cases
Azure Blob Storage	A scalable object storage for **unstructured data** like text, images, videos, and backups. Supports **Hot, Cool, and Archive tiers**.	Media storage, backups, logs, big data processing.
Azure Data Lake Storage (ADLS)	A high-performance storage service optimized for **big data analytics**. Supports **structured, semi-structured, and unstructured data** with hierarchical namespaces.	Data lakes, machine learning, analytics workloads. Gen 2
Azure Files	Fully managed **file shares** in the cloud, accessible via **SMB and NFS protocols**.	File sharing, lift-and-shift applications, application storage.
Azure Queue Storage	A message queue service for **asynchronous communication** between applications, ensuring decoupled architecture.	Message queuing, event-driven architectures, background processing.
Azure Table Storage	A **highly scalable NoSQL key-value store** for structured, non-relational data, enabling **fast lookups** and massive data storage without complex relationships.	Logging events, telemetry data, sensor readings, metadata storage for web apps.

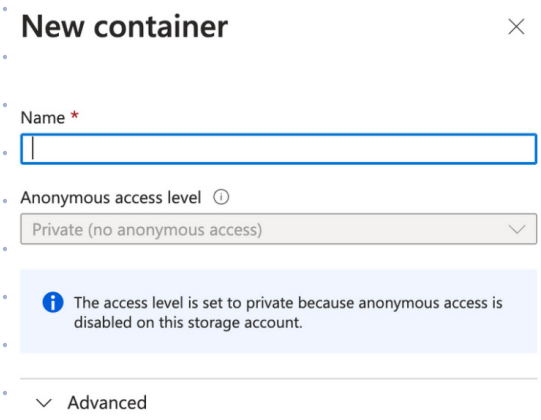
Create Storage Account



) Local
Global

Create Container Inside Storage Account;

Container is used to store Blobs (Unstructured Data)



{ We can change excess level

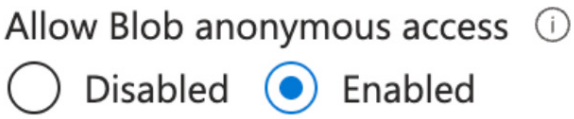
Not preffered.

Once done we will upload Data inside the Container (View Edit)

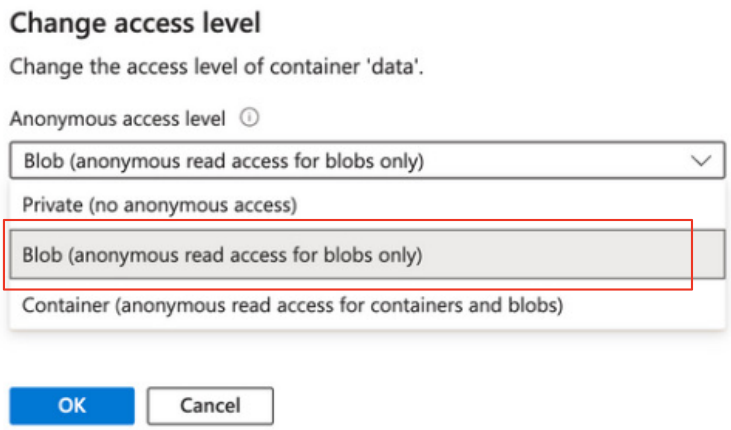
Max 2.1MB Blob (File) can be viewed

Allow Access To Container File

Storage Account → Configuration



Container → File → Change Access Level



Access In Python;

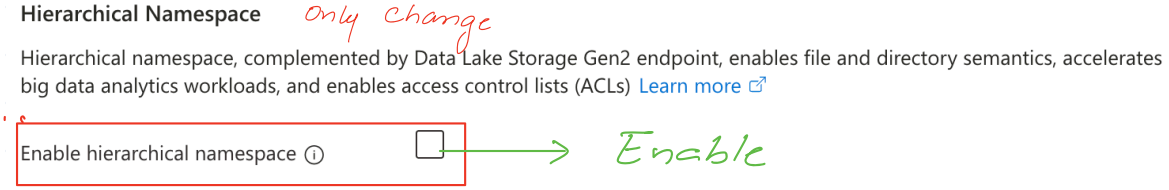
```python
import pandas as pd

url = "https://testingtiku1.blob.core.windows.net/data/ActivityLog-01.csv"
df = pd.read_csv(url)

print(df)
```

Create Gen 2 (Data Lake)

Same Procedure (Hierarchical Namespace)



Create Container Inside Gen2 Storage Account:

- Gen 2 will have more options
- Upload Data

Connect Azure Blob to Power BI

Get Data

↓

Azure

↓

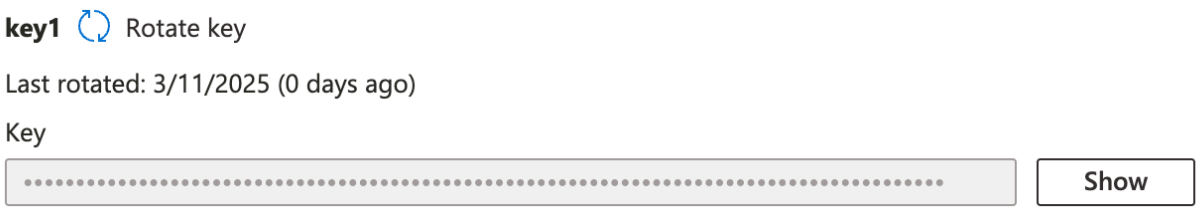
Azure Blob Storage

↓

Enter Storage Account Name

↓

Access Keys (security)



Two ways to Access Files

- Allow Blob anonymous access
  - Disabled
  - Enabled

- Account Name / Access Key