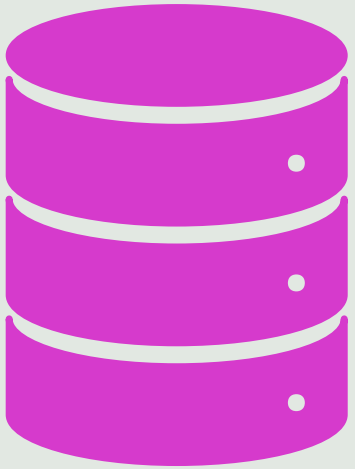
Introduction to Azure Data Lake Storage Gen2

Data Lake Storage Gen2

- Azure Data Lake Storage Gen2 is a set of capabilities dedicated to **big data analytics**, built on Azure Blob Storage
- Data Lake Storage Gen2 converges the capabilities of Azure Data Lake Storage Gen1 with Azure Blob Storage
- For example, Data Lake Storage Gen2 **provides file system semantics**, **file-level security**, and **scale**



Data lake usability

- Typical uses for a data lake include data exploration, data analytics, and machine learning
- With this approach, the raw data is ingested into the data lake and then transformed into a structured queryable format
- Data lake stores are often used in event streaming or IoT scenarios, because they can persist large amounts of relational and nonrelational data without transformation or schema definition

Designed for enterprise big data analytics

- Performance is optimized because you don't need to copy or transform data as a prerequisite for analysis
- Management is easier because you can organize and manipulate files through directories and subdirectories
- Security is enforceable because you can define POSIX permissions on directories or individual files

Key features of Data Lake Storage Gen2

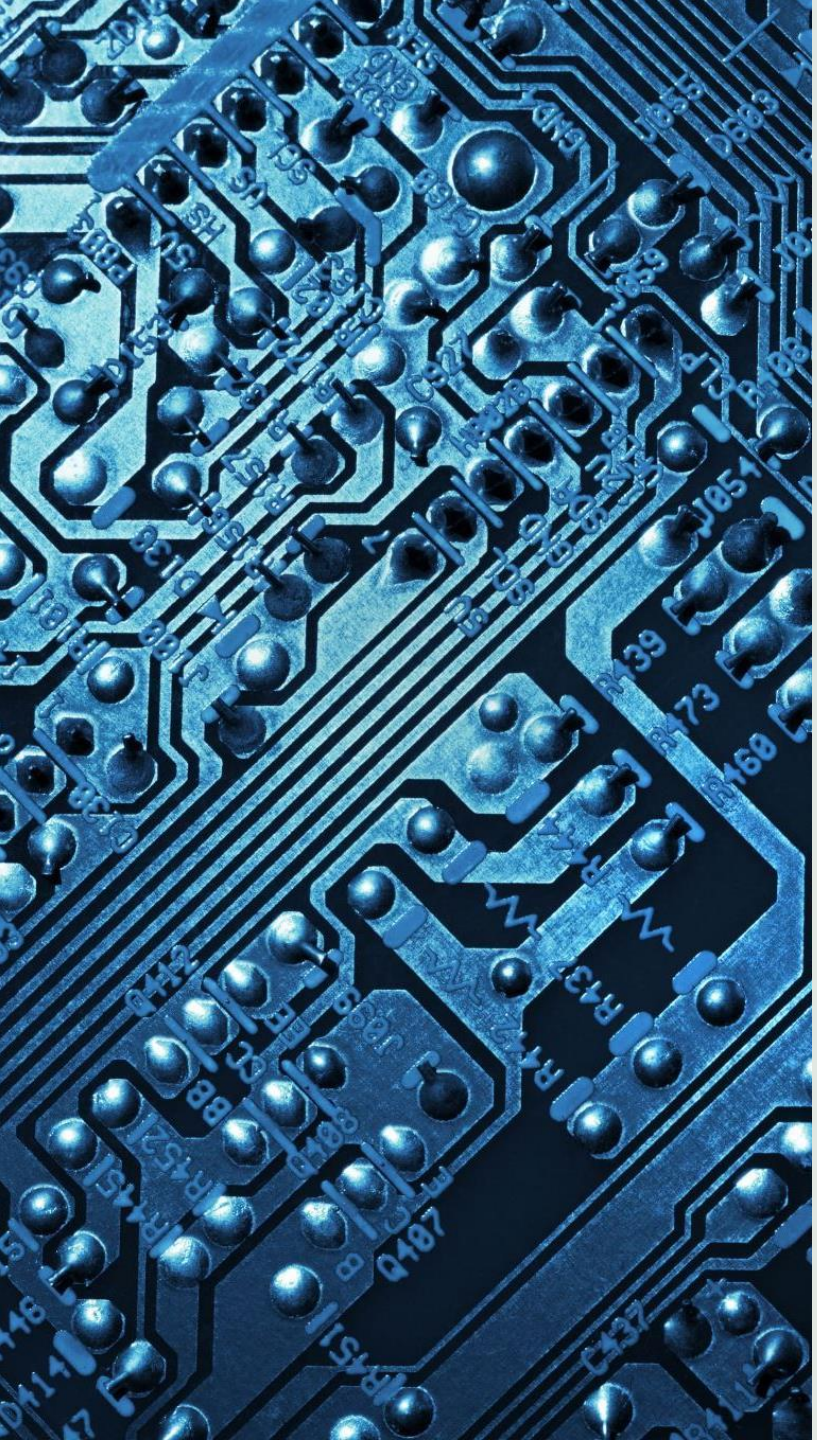
- Hadoop compatible access: Data Lake Storage Gen2 allows you to manage and access data just as you would with a Hadoop Distributed File System
- A superset of POSIX permissions: The security model for Data Lake Gen2 supports ACL(Access Control List) and POSIX permissions along with some extra granularity specific to Data Lake Storage Gen
- Cost-effective: Data Lake Storage Gen2 offers low-cost storage capacity and transactions
- Optimized driver: The **A**zure **B**LOB **F**ile **S**ystem(ABFS) driver is optimized specifically for big data analytics

Scalability

Azure Storage is scalable by design whether you access via Data Lake Storage Gen2 or Blob storage interfaces

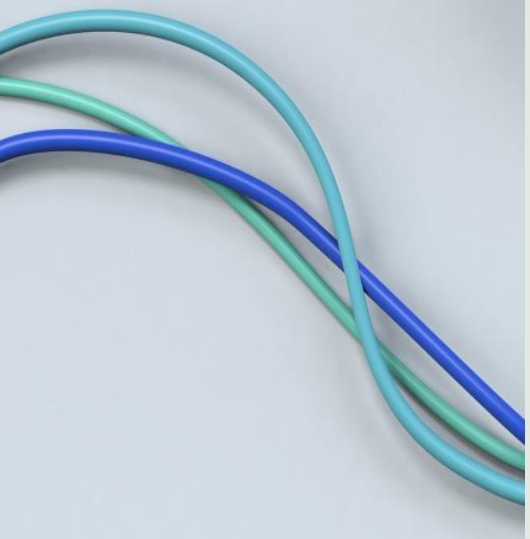
This amount of storage is available with throughput measured in gigabits per second at high levels of input/output operations per second

Processing is executed at near-constant per-request latencies that are measured at the service, account, and file levels



Cost effectiveness

- Because Data Lake Storage Gen2 is built on top of Azure Blob Storage, storage capacity and transaction costs are lower
- Additionally, features such as the hierarchical namespace significantly improve the overall performance of many analytics jobs
- This improvement in performance means that you require less compute power to process the same amount of data, resulting in a lower total cost of ownership for the end-to-end analytics job



One service, multiple concepts

- Because Data Lake Storage Gen2 is built on top of Azure Blob Storage, multiple concepts can describe the same, shared things
- The following are the equivalent entities, as described by different concepts
- Unless specified otherwise these entities are directly synonymous

Supported Blob Storage features

- Blob Storage features such as diagnostic logging, access tiers, and Blob Storage lifecycle management policies are available to your account
- Most Blob Storage features are fully supported, but some features are supported only at the preview level or not yet supported
- To see how each Blob Storage feature is supported with Data Lake Storage Gen2, see Blob Storage feature support in Azure Storage accounts

Supported Azure service integrations

- Data Lake Storage gen2 supports several Azure services
- You can use them to ingest data, perform analytics, and create visual representations
- For a list of supported Azure services, see Azure services that support Azure Data Lake Storage Gen

Supported open source platforms

- Several open source platforms support Data Lake Storage Gen
- Complete list, of Open source platforms that support Azure Data Lake Storage Gen there is in next slide

Open Source Platform

Platform	Supported Version(s)	More Information
HDInsight	3.6+	Apache Hadoop components and versions available with HDInsight
Hadoop	3.2+	Apache Hadoop
Cloudera	6.1+	Cloudera Enterprise 6.x
Azure Databricks	5.1+	Databricks Runtime versions
Hortonworks	3.1.x++	Configuring cloud data access

Technology Choices

- Azure HD Insight is a managed, full-spectrum, open-source analytics service in the cloud for enterprises
- Azure Data Lake Store is a hyperscale, Hadoop-compatible repository
- Azure Data Lake Analytics is an on-demand analytics job service to simplify big data analytics

Challenges

- Lack of a schema or descriptive metadata can make the data hard to consume or query
- Lack of semantic consistency across the data can make it challenging to perform analysis on the data, unless users are highly skilled at data analytics
- It can be hard to guarantee the quality of the data going into the data lake
- Without proper governance, access control and privacy issues can be problems
- A data lake may not be the best way to integrate data that is already relational
- By itself, a data lake does not provide integrated or holistic views across the organization
- A data lake may become a dumping ground for data that is never actually analyzed or mined for insights