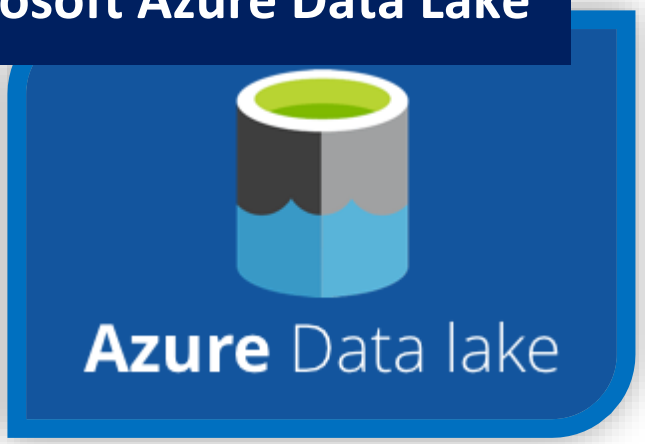
Microsoft Azure Data Lake



Cognitive Convergence is Subject Matter Expert in Office 365, Dynamics 365, SharePoint, Project Server, Power Platform: Power Apps-Power BI-Power Automate-Power Virtual Agents.
We offer Power BI consulting services covering solution architecture refinement, customization, integration, transformation, visualization and analytics to uncover insights hidden within data and enhance data exploration.

CONTENTS

OBJECTIVES

In this paper we will discuss what Azure Data Lake is and what its use cases are, key components, parts and features of Azure Data Lake, and how it works. In this paper Azure Data Lake Storage Gen 1, and Azure Data Lake Storage Gen 2 along with their features are also discussed. Why to use Azure data lake, its benefits, use cases, difference between Data Warehouse and Data Lake and pricing are also discussed. Implementation of creating a storage account with Azure Data Lake storage and analyzing data in Azure Data Lake Storage Gen 3 by using Power BI is covered here.

MICROSOFT AZURE DATA LAKE

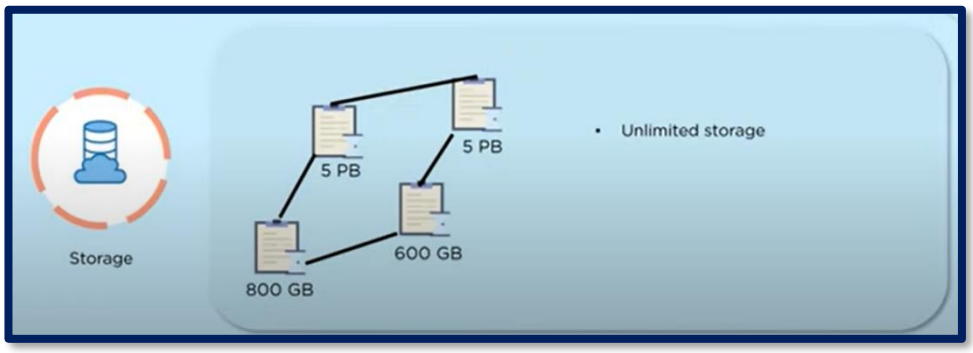
Azure Data Lake is a highly scalable, distributed, parallel file system in the cloud that is specifically designed to work with multiple analytics frameworks.

The data in output datasets (collected from mobile, the web, social platforms, etc.) is sent into the Azure Data Lake Store. It is then provided to external frameworks, like R and Apache Spark.

Data Lake works on two main concepts: storage and analytics.

Storage

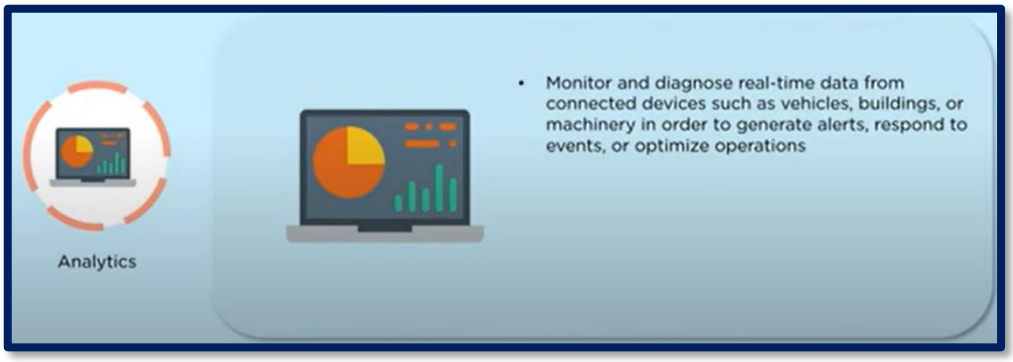
Storage is unlimited, allowing users to save very large files. A variety of data (like unstructured or structured data) can be stored here.



Analytics

Through analytics, you can monitor and diagnose real-time data from connected devices, such as vehicles, buildings, or machinery to initiate actions such as generating alerts, responding to events, and optimizing operations. Financials can also be monitor

- Financial transactions in real-time to detect fraudulent activity
- The use of a credit card across geographic locations
- The number of transactions on a single credit card



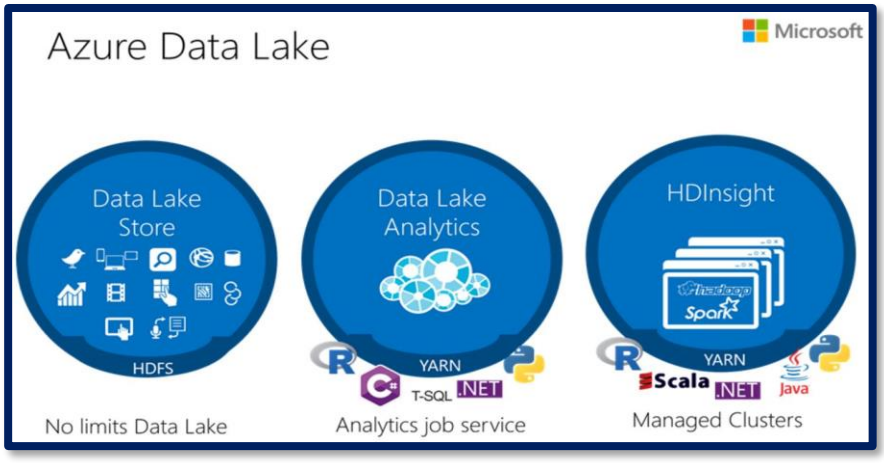
WHY AZURE DATA LAKE

- Machine Learning and Artificial Intelligence can be used to make profitable prediction.
- Data Lake give 360 degree view of customers and makes analysis more robust.

HOW DOES AZURE DATA LAKE WORK?

Azure Data Lake is built on Azure Blob storage, which is the Microsoft object storage solution for the cloud. The solution features low-cost, tiered storage and high-availability/disaster recovery capabilities. It integrates with other Azure services, including Azure Data Factory, which is a tool for creating and running extract, transform and load (ETL) and extract, load and transform (ELT) processes.

The solution is based on the Apache Hadoop YARN (Yet Another Resource Negotiator) cluster management platform. It can scale dynamically across SQL servers within the data lake, as well as servers in Azure SQL Database and Azure SQL Data Warehouse.

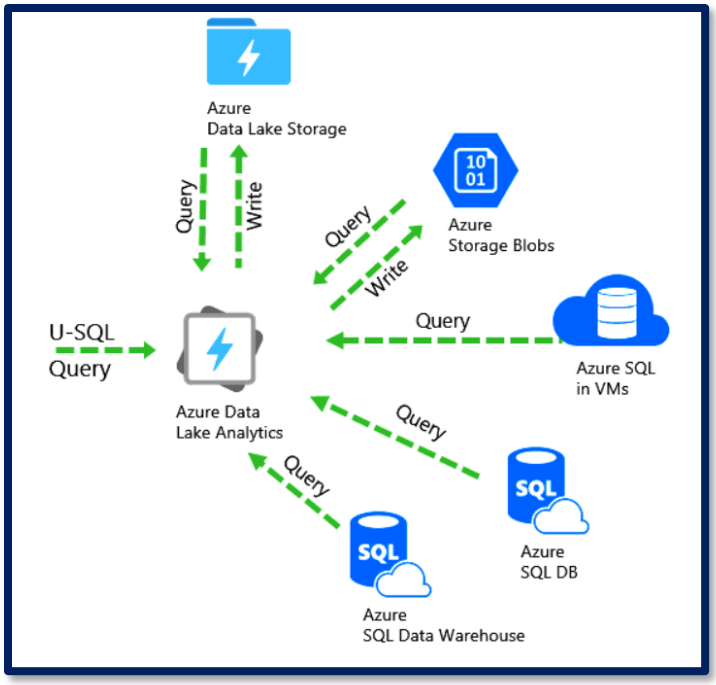


THREE PARTS OF AZURE DATA LAKE

The full solution consists of three components that provide storage, an analytics service and cluster capabilities.

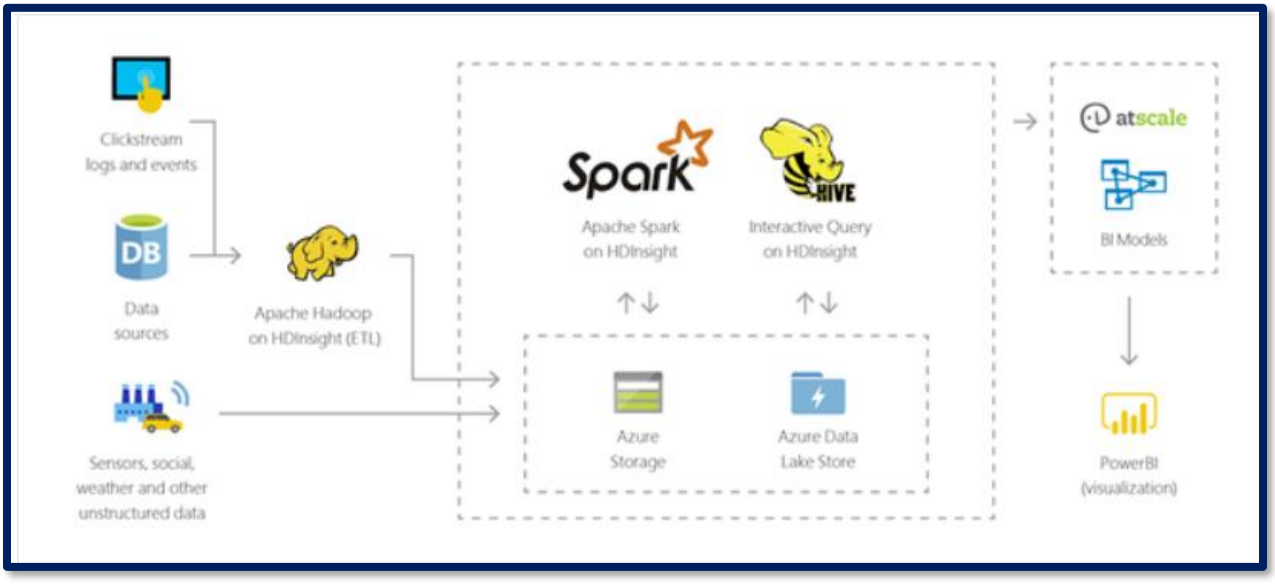
Azure Data Lake Analytics

Azure Data Lake Analytics is an on-demand analytics job service built on Apache Hadoop YARN that simplifies big data. It processes big data jobs in seconds and no infrastructure to worry about because there are no virtual machines, servers, or clusters to wait for, manage, or tune. It is designed to let users perform analytics on data up to petabytes in size. It covers U-SQL, a query language that extends the simple, familiar, declarative nature of SQL with the dramatic power of C#. It is a cost-effective solution for big data workloads. You pay on a per-job basis when data is processed.



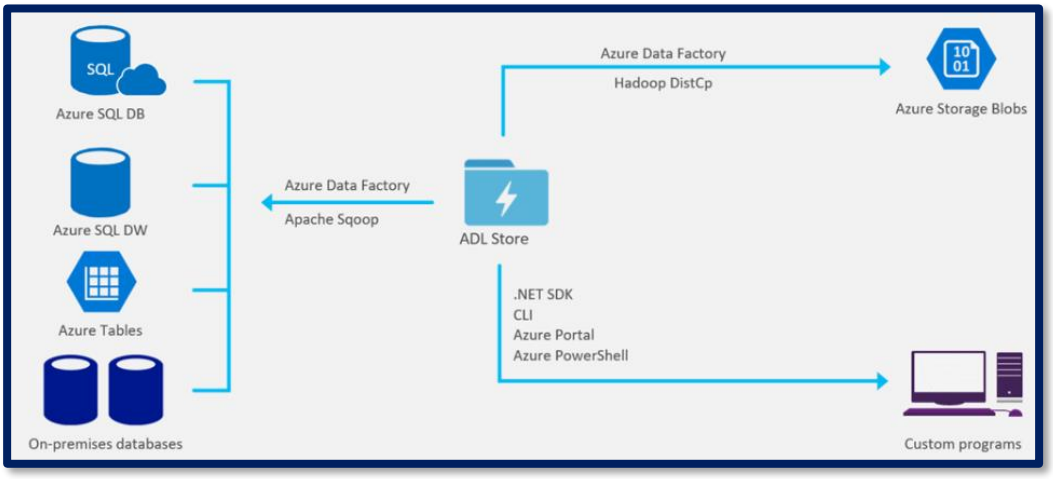
Azure HDInsight

Azure HDInsight is a cluster management solution that makes it easy, fast, and cost-effective to process massive amounts of data. It's a cloud deployment of Apache Hadoop that enables users to take advantage of optimized open source analytic clusters for Apache Spark, Hive, Map Reduce, HBase, Storm, Kafka, and R-Server. With these frameworks, you can support a broad range of functions, such as ETL, data warehousing, machine learning, and IoT. Azure HDInsight also integrates with Azure Active Directory for role-based access controls and single sign-on capabilities.



Azure Data Lake Store

Azure Data Lake Storage is a massively scalable and secure data lake for high-performance analytics workloads. Azure Lake Data Storage was formerly known and is sometimes still referred to as the Azure Data Lake Store. Designed to eliminate data silos, Azure Data Lake Storage provides a single storage platform that organizations can use to integrate their data. Azure Data Lake Storage can help optimize costs with tiered storage and policy management. It also provides role-based access controls and single sign-on capabilities through Azure Active Directory. Users can manage and access data within Azure Data Lake Storage using the Hadoop Distributed File System (HDFS). Therefore any tool that you're already using that is based on HDFS will work with Azure Data Lake Storage.



AZURE DATA LAKE STORAGE GEN 1

Azure Data Lake Storage Gen1 is an enterprise-wide hyper-scale storehouse for big-data analytic workloads. It permits us to capture data of any type, size, and ingestion speed in one single place for operational and exploratory analytics. It carries all enterprise-grade capabilities such as scalability, security, manageability, availability, and reliability.

Big data analytics

ADLS Gen1is built for running large-scale analytic systems that require huge throughput to analyze and query large amounts of data.

KEY FEATURES OF DATA LAKE STORAGE GEN1

Some of the key features of Data Lake Storage Gen1 include the following.

Made for Hadoop

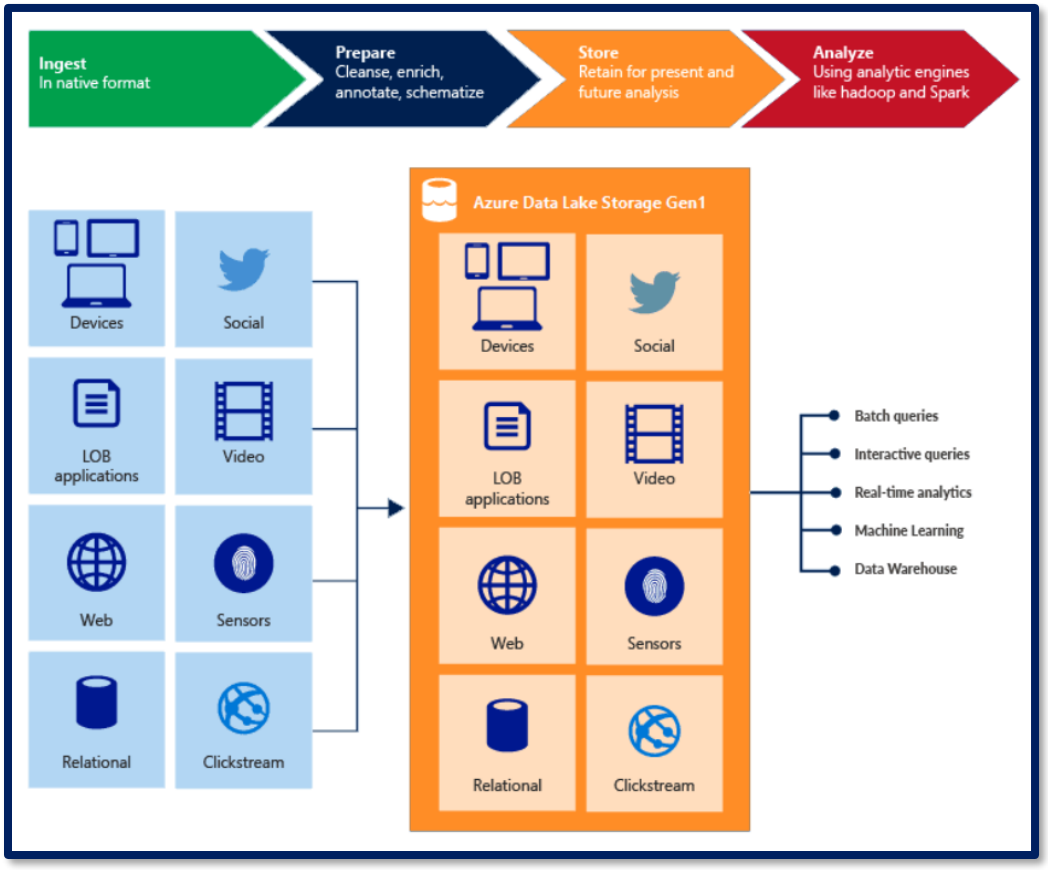
One can easily analyze data stored in ADLS Gen1 using Hadoop analytic frameworks such as Hive or MapReduce.

Unlimited storage

ADLS Gen1provides unlimited storage and can store a range of data for analytics and range from kilobytes to petabytes in size.

Highly available and securing data

In ADLS Gen1 data are stored securely by making redundant copies to guard against any sudden failures.



AZURE DATA LAKE STORAGE GEN2

- ADLS Gen2 is a collection of capabilities for big data analytics.
- It is built on Azure Blob storage, and have all the key features of ADLS Gen1.
- ADLS Gen2 offers capabilities like:
  - file system semantics
  - file-level security
  - directory
  - low-cost
  - scalability
  - high availability/disaster recovery

KEY FEATURES OF DATA LAKE STORAGE GEN2

Some of the key features of Data Lake Storage Gen2 include the following.

Hadoop suitable access

ADLS Gen2 permits you to access and manage data just as you would with a Hadoop Distributed File System (HDFS).

POSIX permissions

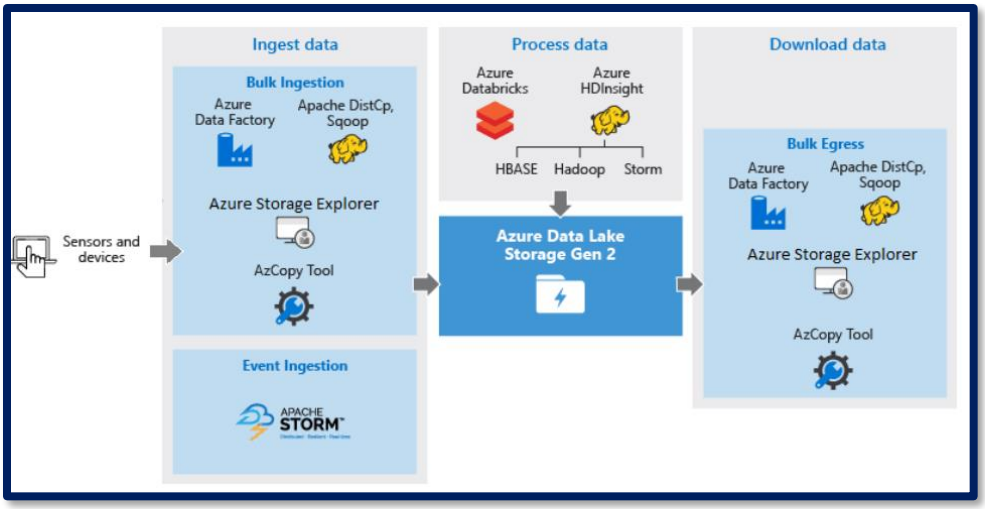
The security design for ADLS Gen2 supports ACL and POSIX permissions along with some more granularity specific to ADLS Gen2.

Low Cost

ADLS Gen2 offers low-cost transactions and storage capacity.

Optimized driver

The ABFS driver is developed exactly for big data analytics.



USES-CASES OF AZURE DATA LAKE

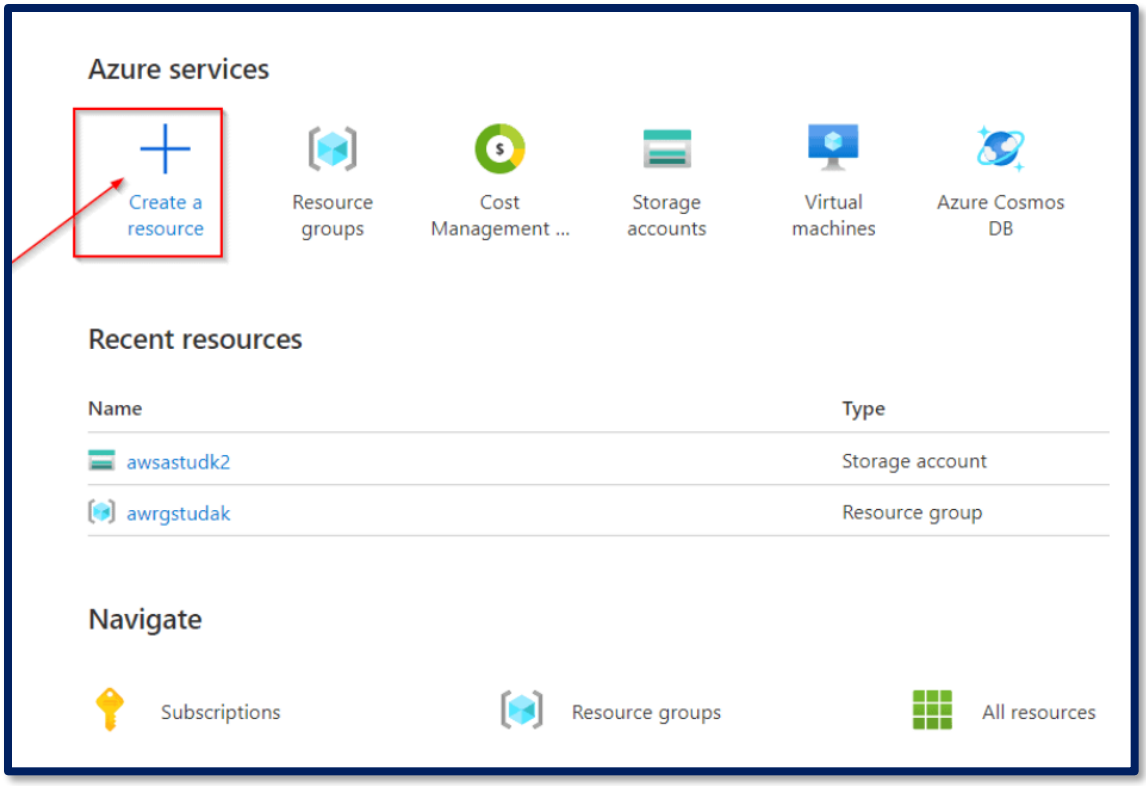
- General-purpose object storage handled by Azure.
- Streaming and processing of batch workloads.
- Selection of data by analysts and data engineers for specific needs without making copies.

ADVANTAGE OF AZURE DATA LAKE

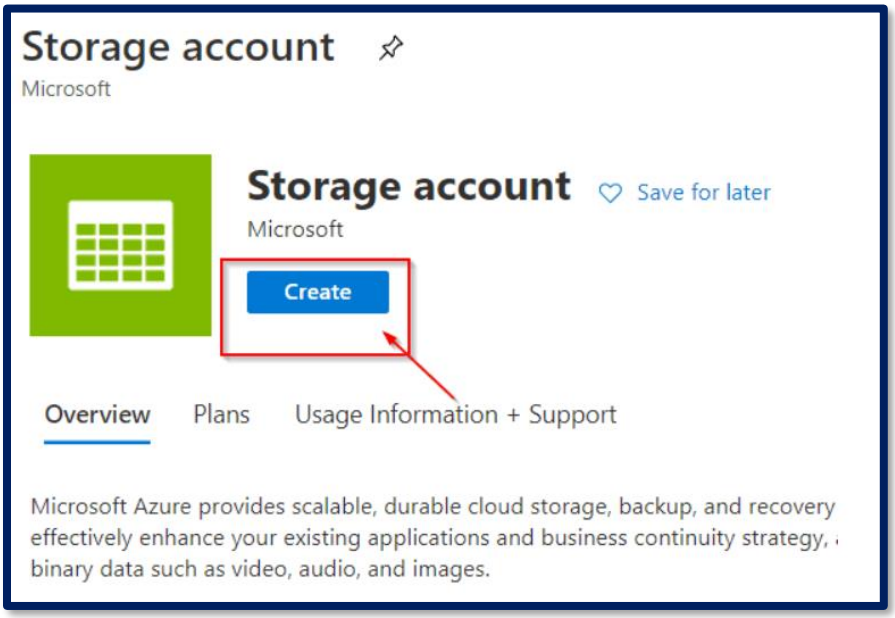
- Highly flexible and scalable as it is housed on the cloud.
- Allows streamlining data storage for all business needs.
- A huge amount of data can be processed simultaneously providing quick access to insights.
- Data Lake stores everything like multimedia, logs, XML, sensor data, social data, binary, chat, and people data.
- No limit on data storage and file size.
- Supports massive analytics workloads for in-depth analytics.
- It supports schema-less storage.

CREATE A STORAGE ACCOUNT WITH AZURE DATA LAKE STORAGE GEN2

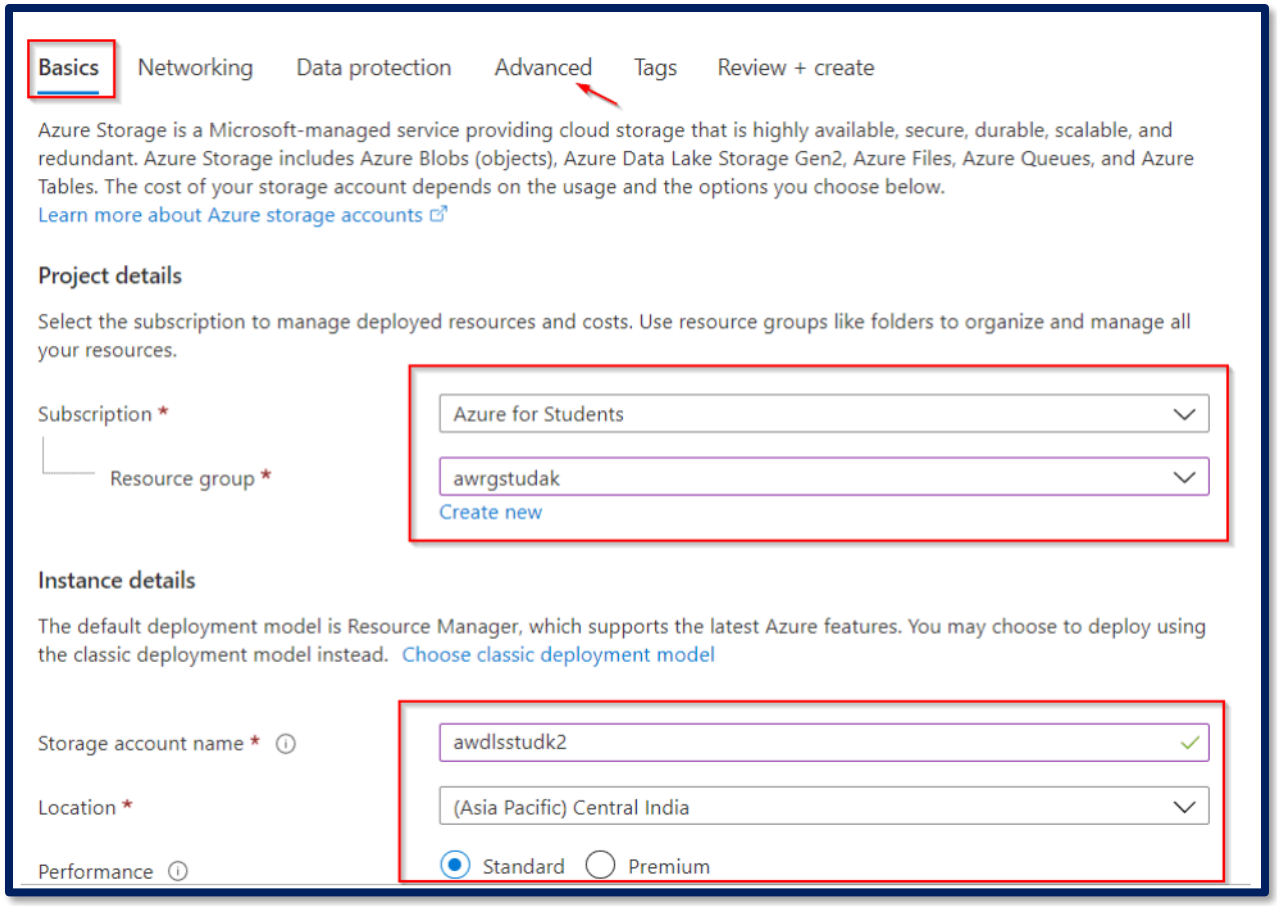
In the Azure portal, click on + Create a resource icon.



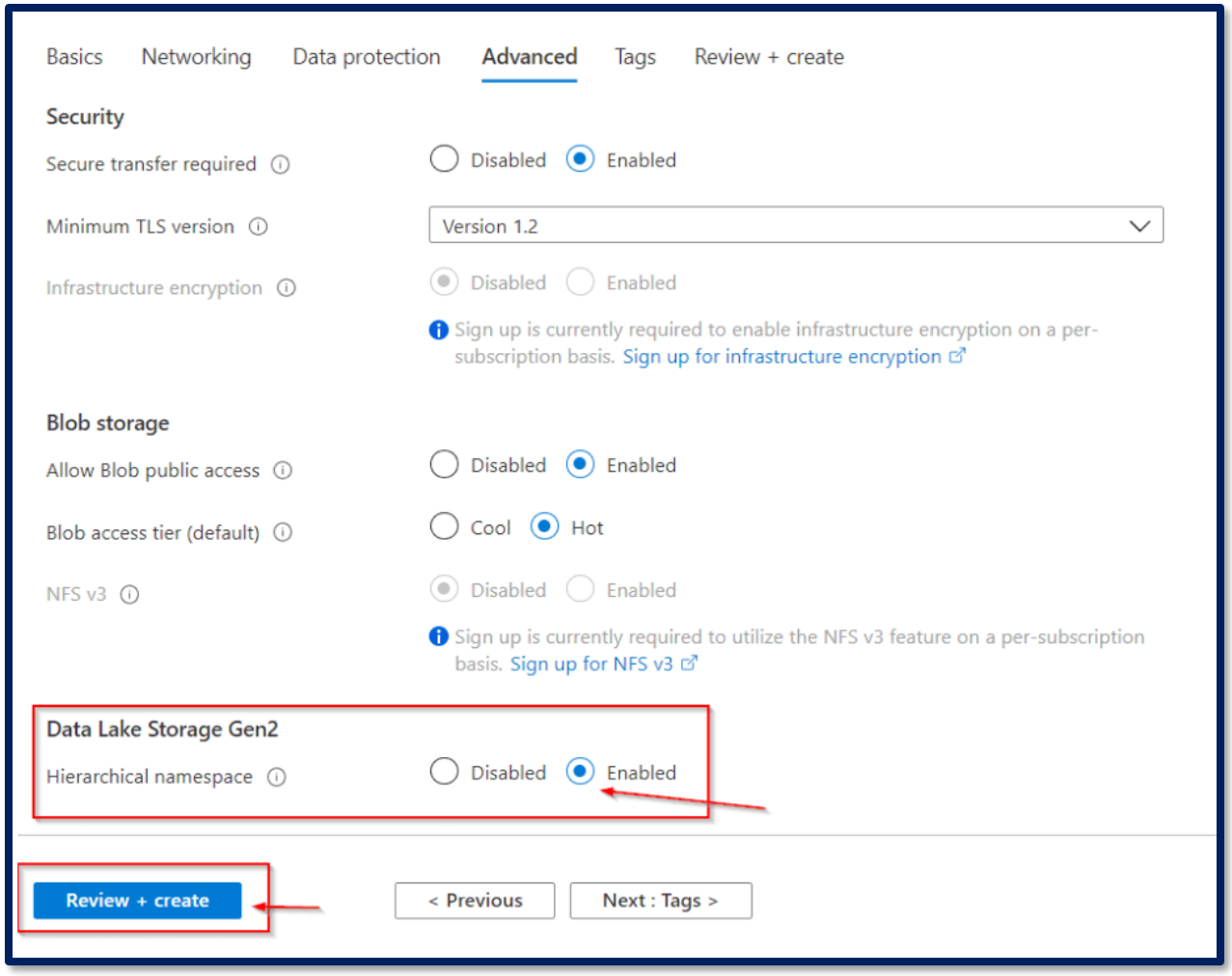
Click in the Search the Marketplace text box, and type the word storage. Click on Storage account in the list that appears. Click Create



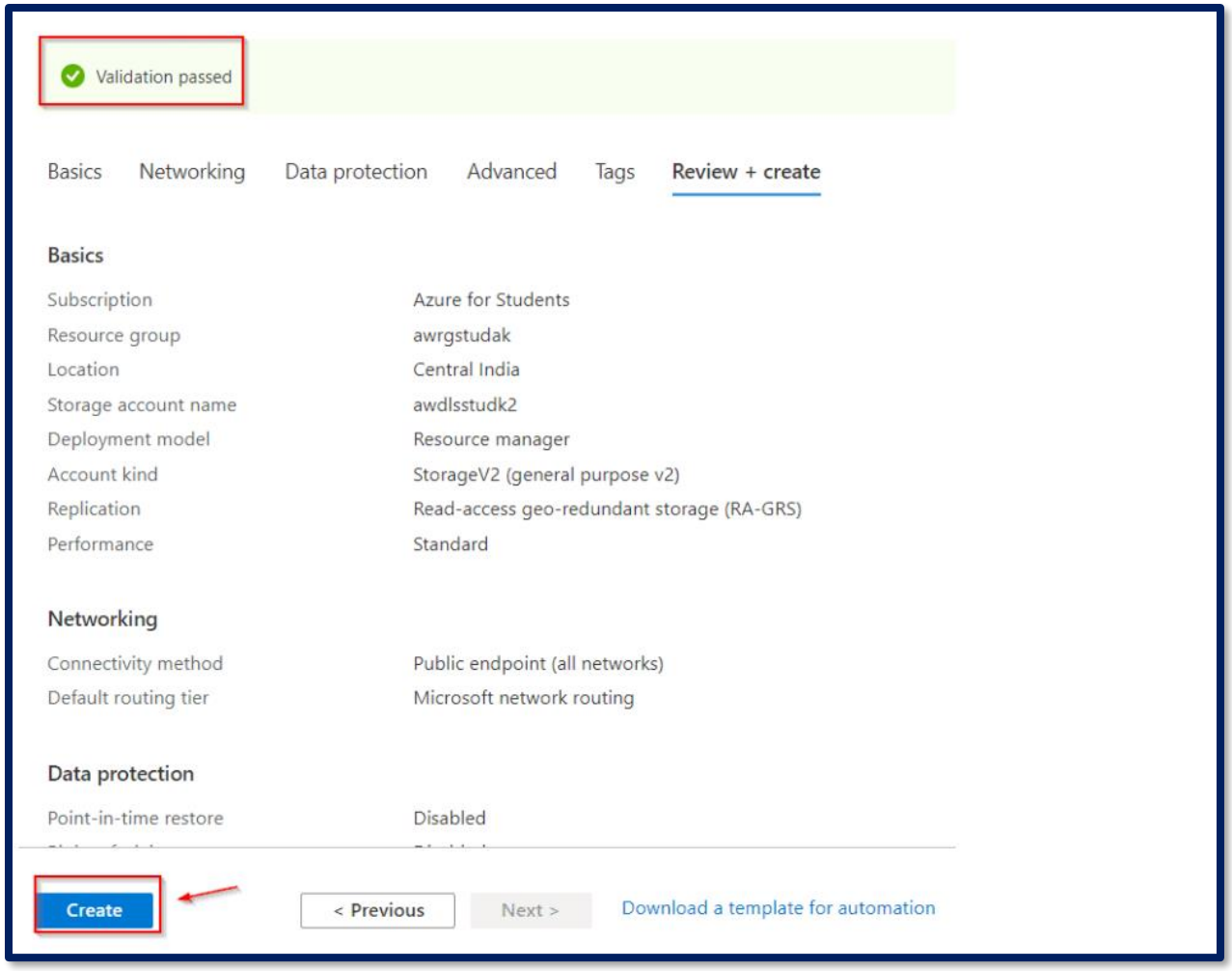
Add project details, storage account detail and location



On the Advanced tab page. Click Enabled under Hierarchical namespace. Then click Review + create.



After the validation of the Create storage account blade, click Create.



HOW TO GET DATA INTO THE DATA LAKE?

In Azure, the most prominent tool for moving data is Azure Data Factory (ADF). ADF is designed to Azure Data Factory Icon move large volumes of data from one location to another making it a key component in your effort to collect data into your Data Lake. V2 of Azure Data Factory was recently released. V2 provides an improved UI, trigger-based execution, and Git integration for building data movement pipelines. Another key addition is support for SSIS package execution so you can reuse existing investments in data movement and transformation.

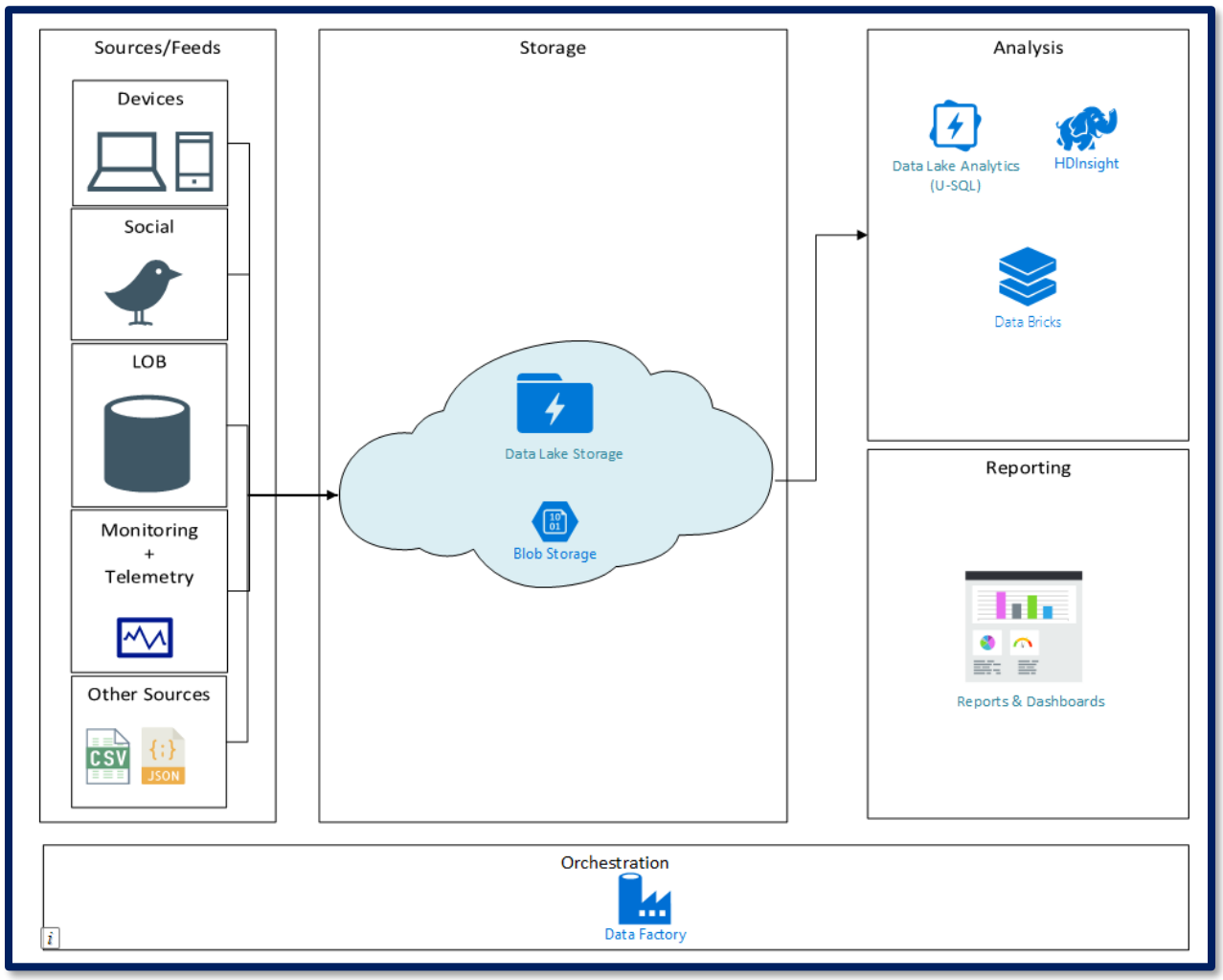
HOW TO PERFORM BATCH ANALYSIS OF DATA IN THE DATA LAKE?

In addition to ADLS, there are other analysis services that are directly applicable to the large-scale batch analysis of unstructured data that resides in your data lake. Two of these services available on Azure are HDInsight and Databricks. Using these other services may make sense if you are already familiar with them and/or they are already part of your

analytics platform in Azure. Databricks provides an Apache Spark SaaS offering that allows you to collaborate and run analytics processes on demand. HDInsight provides a greater range of analytics engines including HBase, Spark, Hive, and Kafka. However, HDInsight is provided as a PaaS offering and therefore requires more management and setup.

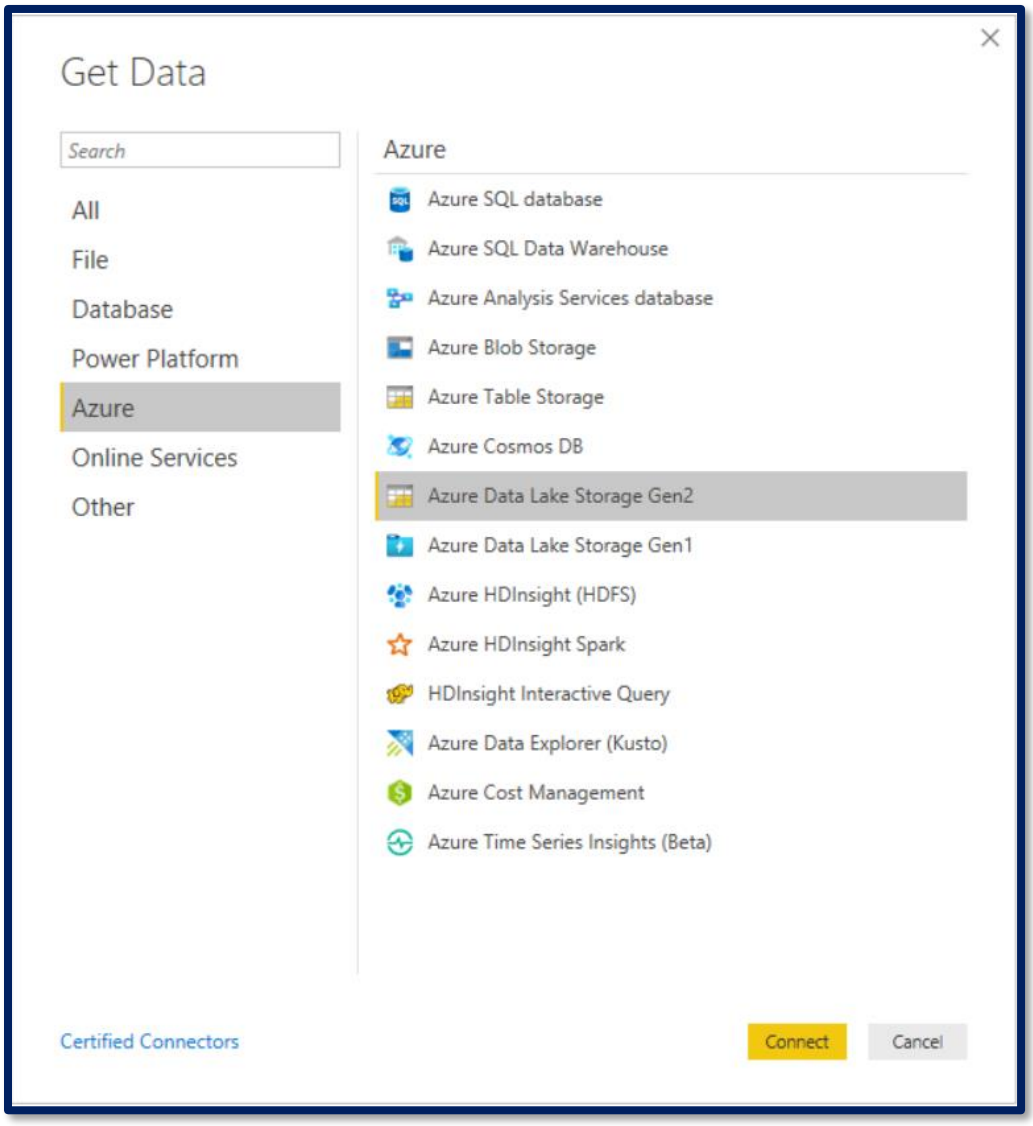
HOW TO REPORT ON DATA IN THE DATA LAKE?

Azure Data Lake and the related tools mentioned above provide the ability to analyze your data but are not the generally the right source for reports and dashboards. Once analyzed data and identified measures and metrics might want to see in dashboards and reports, do some additional work. Ideally, data for dashboards and reports will be structured and stored in a service designed to be queried regularly and update the report or dashboard data. The right place for this data will be a destination like SQL Azure, a SQL Azure Data Warehouse, Cosmos DB or your existing BI platform. This is another stage where Azure Data Factory will be key, as it can orchestrate the process to read data, schedule execution of analysis (if needed), structure data, and write the resulting data to your Reporting data store.



ANALYZE DATA IN AZURE DATA LAKE STORAGE GEN2 BY USING POWER BI

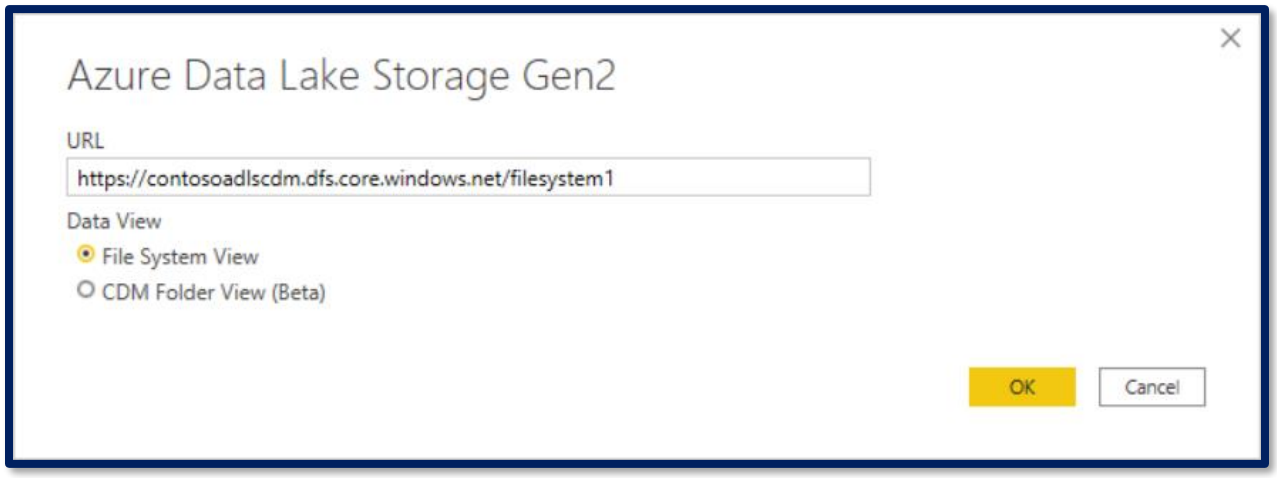
1. Launch Power BI Desktop on your computer
2. From the Home tab of the Ribbon, select Get Data, and then select More.
3. In the Get Data dialog box, select Azure > Azure Data Lake Store Gen2, and then select Connect.



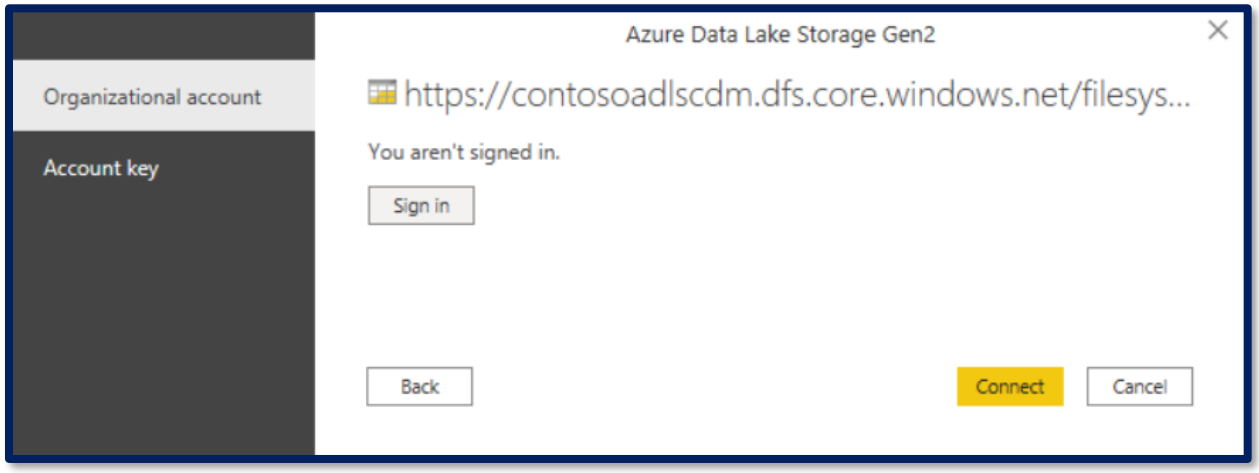
In the Azure Data Lake Storage Gen2 dialog box, provide the URL to Azure Data Lake Storage Gen2 account, filesystem, or subfolder using the container endpoint format. URLs for Data Lake Storage Gen2 have the following pattern:

https://<accountname>.dfs.core.windows.net/<filesystemname>/<subfolder>

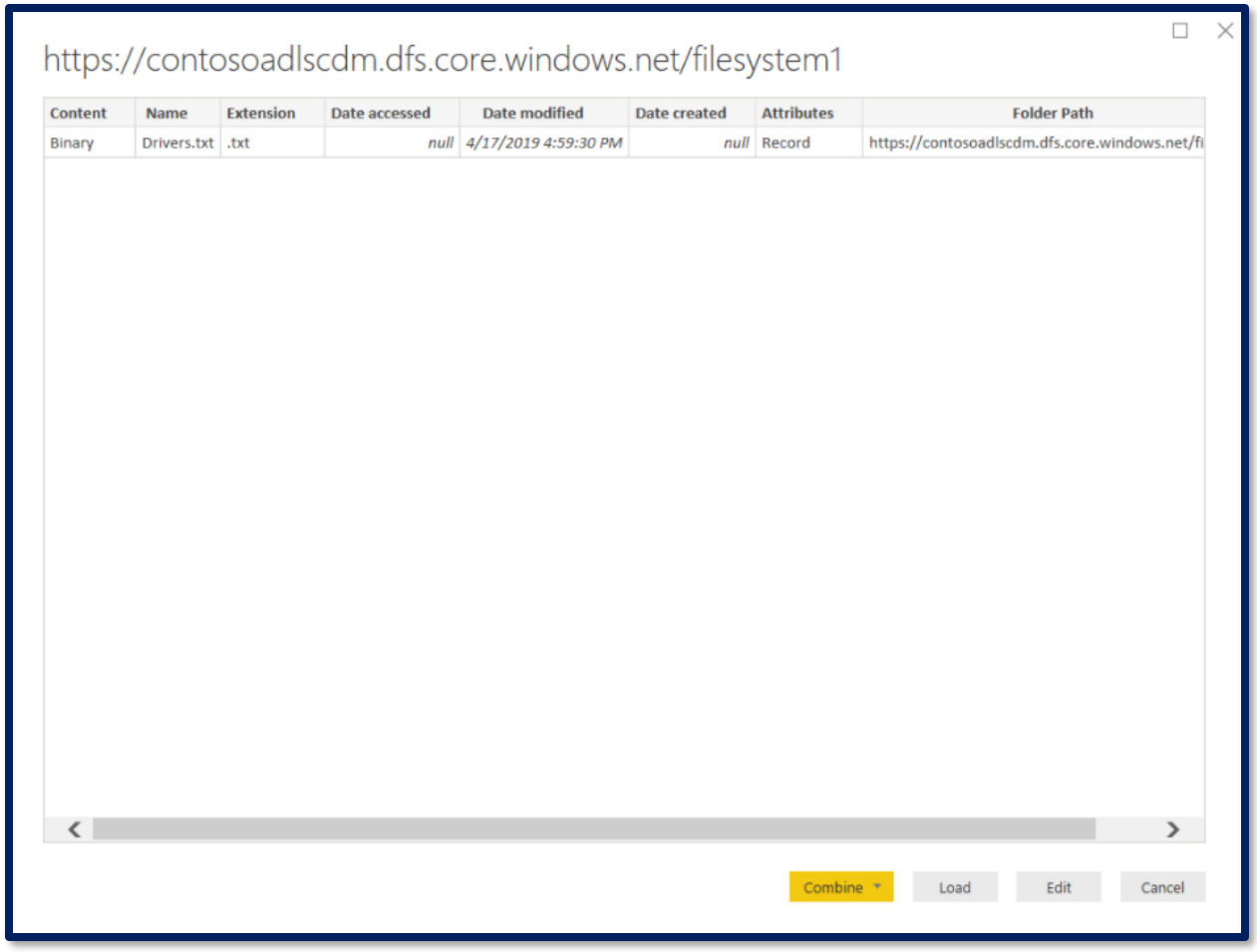
- Select whether to use the file system view or the Common Data Model folder view.
- Select OK to continue.



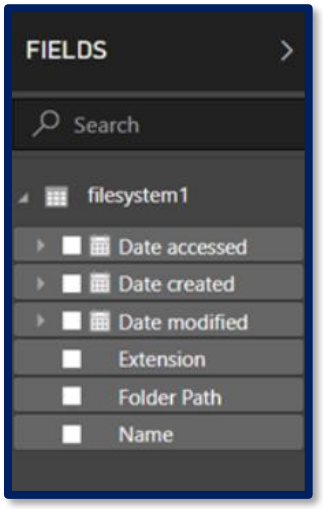
Select the authentication method.



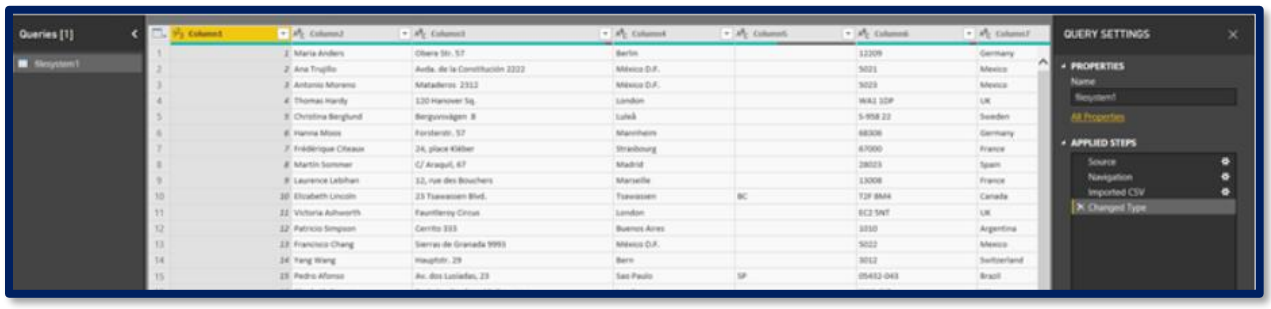
The next dialog box shows all files under the URL, including the file that uploaded to storage account. Verify the information, and then select Load.



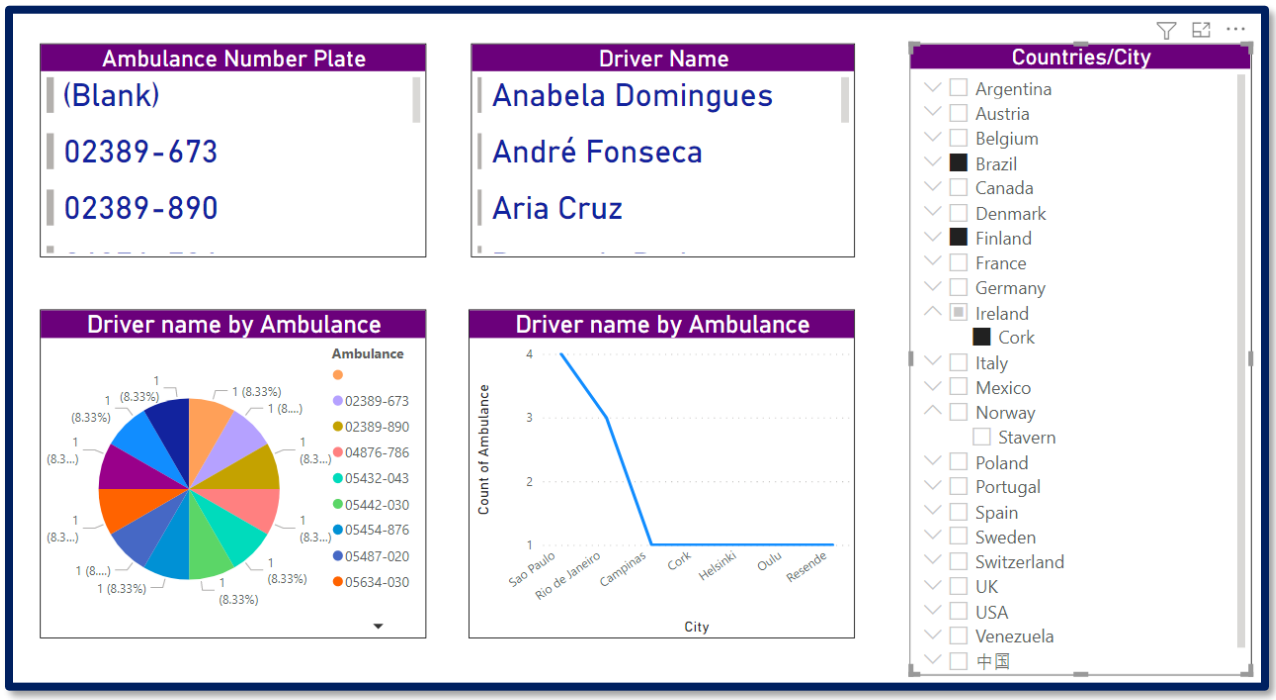
After the data has been successfully loaded into Power BI.



However, to visualize and analyze the data, the data to be available using the following fields. In the Query Editor, under the Content column, select Binary. The file will automatically be detected as CSV and allow to see an output as shown below. Data is now available in a format that can use to create visualizations.



Once the query is updated, the Fields tab will show the new fields available for visualization. Now visualization can created.



DIFFERENCE BETWEEN DATA WAREHOUSE AND DATA LAKE

	Data Warehouse	Data Lake
Data	Structured and Processed	Semi-structured, unstructured and Structured
Processing	Schema on write	Schema on reading
Storage	Expensive	Low cost
Agility	Less agile and fixed-configuration	Highly agile and fully configurable
Security	Mature	Mature
Users	Business professionals	Data Scientists

PRICING

Pricing for Azure Data Lake is dependent upon numerous variables, including storage capacity, the number of analytics units (AUs) per minute, the number of completed jobs and the cost of managed Hadoop and Spark clusters. As of this writing, the Azure Data Lake Store service is priced at $0.039 per GB per month for pay as you go, with capacity-based discounts up to 33% for monthly commitments. The Azure Pricing Calculator can help customers determine exact data lake costs.

Data Lake Price/Month (Pay-as-you-go)

- First 100 TB: Rs. 2.58 per GB
- Next 100 TB to 1,000 TB: Rs. 2.52 per GB
- Next 1,000 TB to 5,000 TB: Rs. 2.45 per GB

CONCLUSION

In this paper we discussed what Azure Data Lake is and what its use cases are, key components, parts and features of Azure Data Lake, and how it works. Azure Data Lake Storage Gen 1, and Azure Data Lake Storage Gen 2 along with their features, why to use Azure Data Lake, its benefits, use cases, difference between Data Warehouse and Data Lake and pricing are also discussed. Implementation of creating a storage account with Azure Data Lake storage and analyzing data in Azure Data Lake Storage Gen 3 by using Power BI is covered here.

CONTACT US

Shahzad Sarwar
Entrepreneur/Architect/Consultant
Cognitive Convergence
http://www.cognitiveconvergence.com
Voice: +1 4242530744
Skype: Shahzad.Sarwar.Online
shahzad@cognitiveconvergence.com