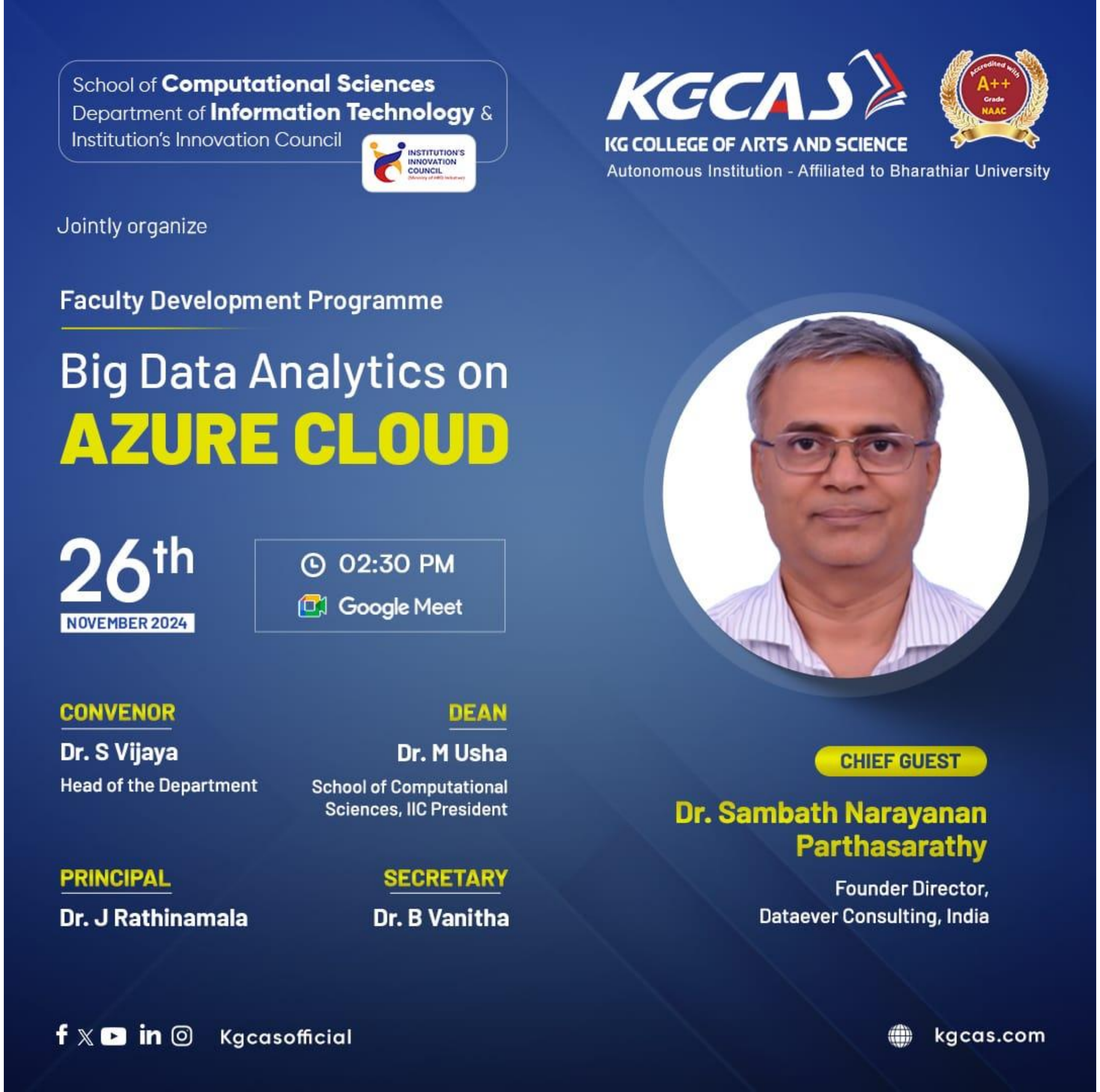
Bigdata Analytics on Azure cloud



School of **Computational Sciences**
Department of **Information Technology** &
Institution's Innovation Council

INSTITUTION'S INNOVATION COUNCIL

KGCAS
KG COLLEGE OF ARTS AND SCIENCE
Autonomous Institution - Affiliated to Bharathiar University

Accredited with A++ Grade NAAC

Jointly organize

Faculty Development Programme

Big Data Analytics on AZURE CLOUD

26th
NOVEMBER 2024

02:30 PM
Google Meet

CONVENOR
Dr. S Vijaya
Head of the Department

DEAN
Dr. M Usha
School of Computational Sciences, IIC President

PRINCIPAL
Dr. J Rathinamala

SECRETARY
Dr. B Vanitha

CHIEF GUEST
Dr. Sambath Narayanan Parthasarathy
Founder Director,
Dataever Consulting, India

Kgcasofficial

kgcas.com

Agenda

1. Data challenges faced by the industries
2. Business Benefits of Data Lake
3. What is Big Data and Data Lake ?
4. Why Cloud is required for Big Data Solution
5. Data Engineering Solution Description
6. Azure Data Engineering End to End Solution Architecture
7. Azure Solution Demo
8. Dashboard

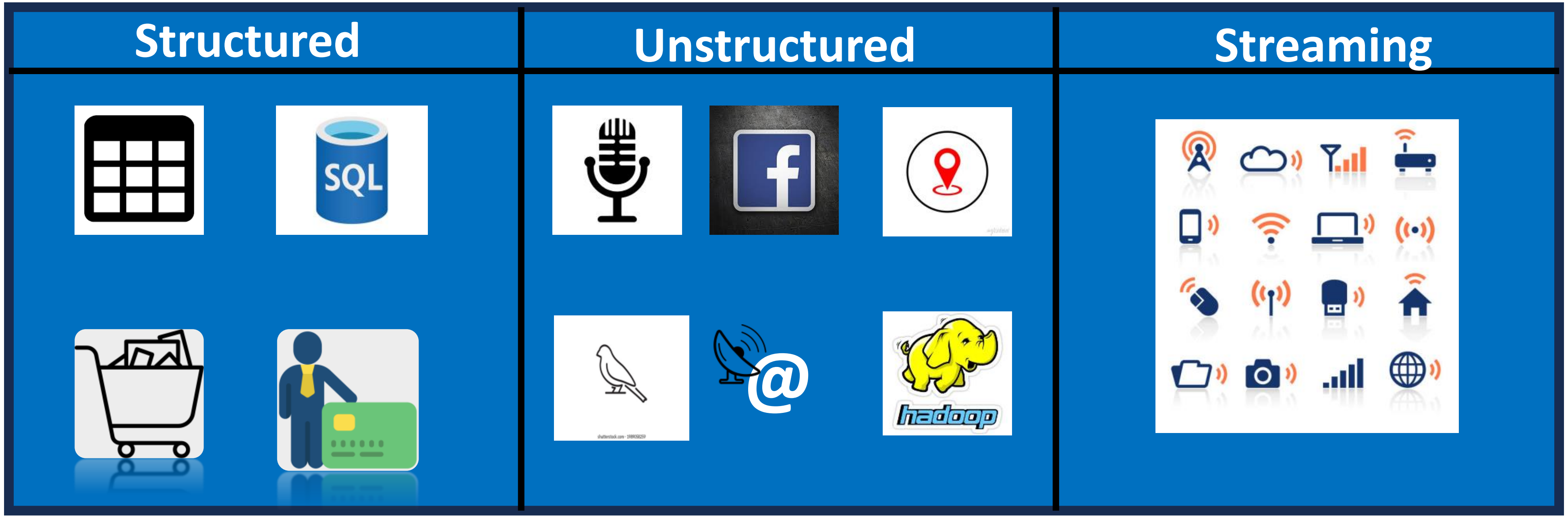
Objective

1. Build one common Industrial Data Lake solution on Azure cloud
2. End to End Project
3. The primary goal of this project is to build industrial end-to-end Data Lakes on Microsoft Azure cloud.
4. For this purpose, use an opensource dataset.
5. A data lake solution usually comprises of a storage layer, a compute layer, and a serving layer. The compute layers on cloud include Extract, Transform, Load (ETL); Batch; or Stream processing.

Problem faced by industries

Source Data are in different format and are in different location.

- Email , Cloud , SQL , Excel , Pdf ,Doc ,Image ,Real time data on AWS ,Text



Structured

Unstructured

Streaming

Sample Source Datasets in open domain

We have used these datasets in our projects

1. AdventureWorksLT2019 is a SQL Dataset
2. Real time data collected from connected car is in Json in **AWS S3.**
3. Motorcycle data set in csv format from Kaggle.
4. Inventory Management Sap Hana Data base from Sap BTP.
5. Car insurance data set in csv format from Kaggle.

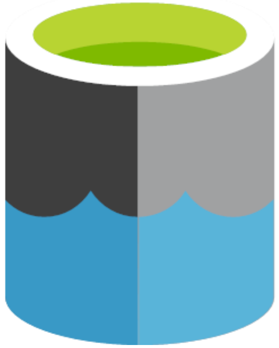
Business Benefits of Data Lake

1. From Different Data in different format, store them in data lake and extract business insights from the data.
2. Accessibility , security , data protection and disaster recovery are simplified

What is Big Data and Data Lake ?

Big data refers to extremely large and diverse collections of **structured, unstructured, and semi-structured** data that continues to grow exponentially over time. These datasets are so huge and complex in **volume, velocity, and variety**.

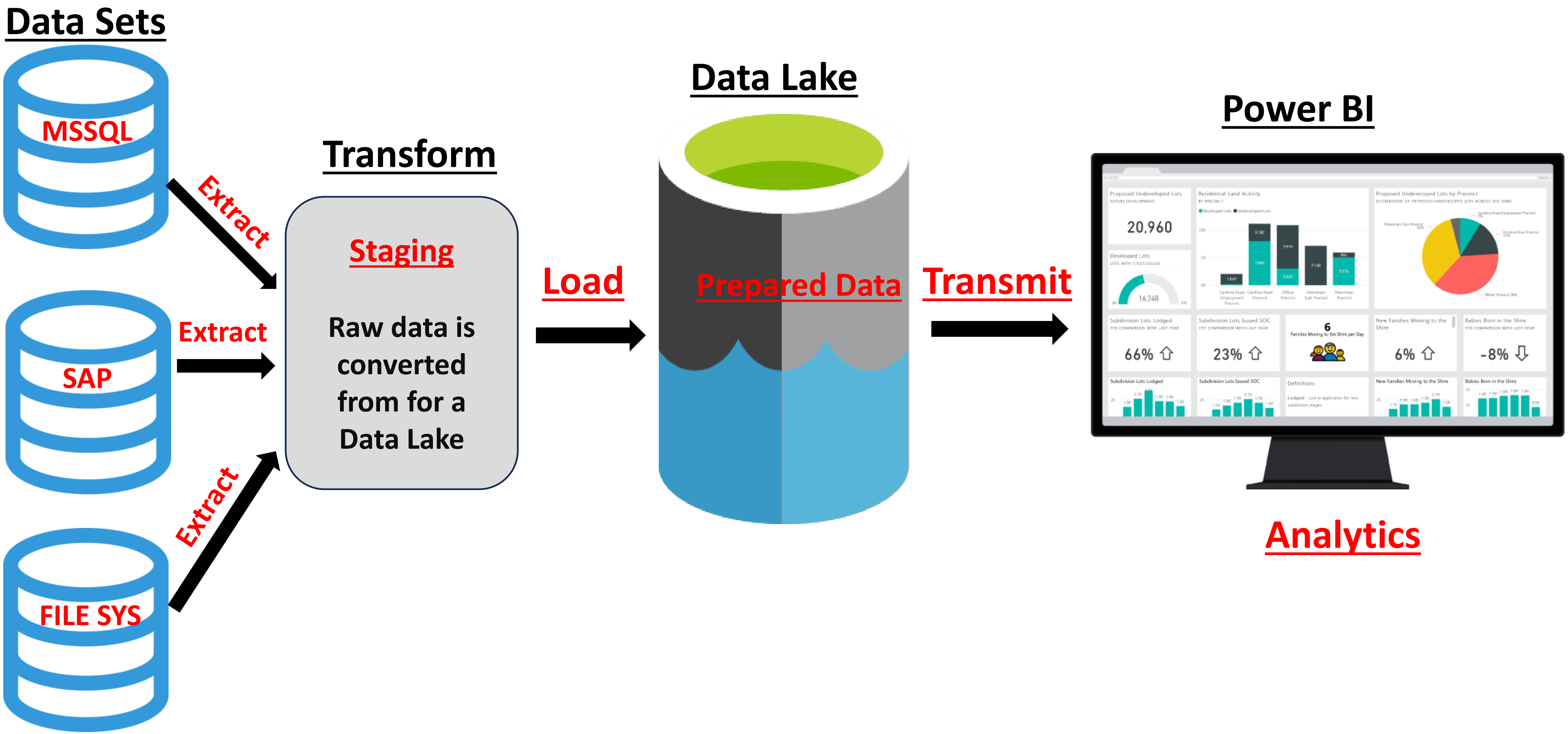
Difference Between Data Lake and Data Warehouse

Data Lake	Data Warehouse
A data lake holds raw and unstructured data, in one location. which is then ready to be used across applications  Data Lake	A data warehouse is a system that pulls structured, pre-defined data from a variety of sources and processes that data for operational use.

ETL (Extract, Transform, Load) process

1. **Extract, Transform, Load (ETL)** is the general procedure of copying data from one or more data sources into a destination system which represents the data differently from the source(s). The ETL process is often used in data warehousing and data lake.
2. **Data extraction** involves extracting data from homogeneous or heterogeneous sources
3. **Data transformation** processes data by cleaning and transforming them into a proper storage format/structure for the purposes of querying and analysis
4. **Data loading** describes the insertion of data into the target data store, data mart, data lake or data warehouse.

ETL (Extract, Transform, Load)



Data Sets

MSSQL

SAP

FILE SYS

Extract

Extract

Extract

Transform

Staging

Raw data is converted from for a Data Lake

Load

Data Lake

Prepared Data

Transmit

Power BI

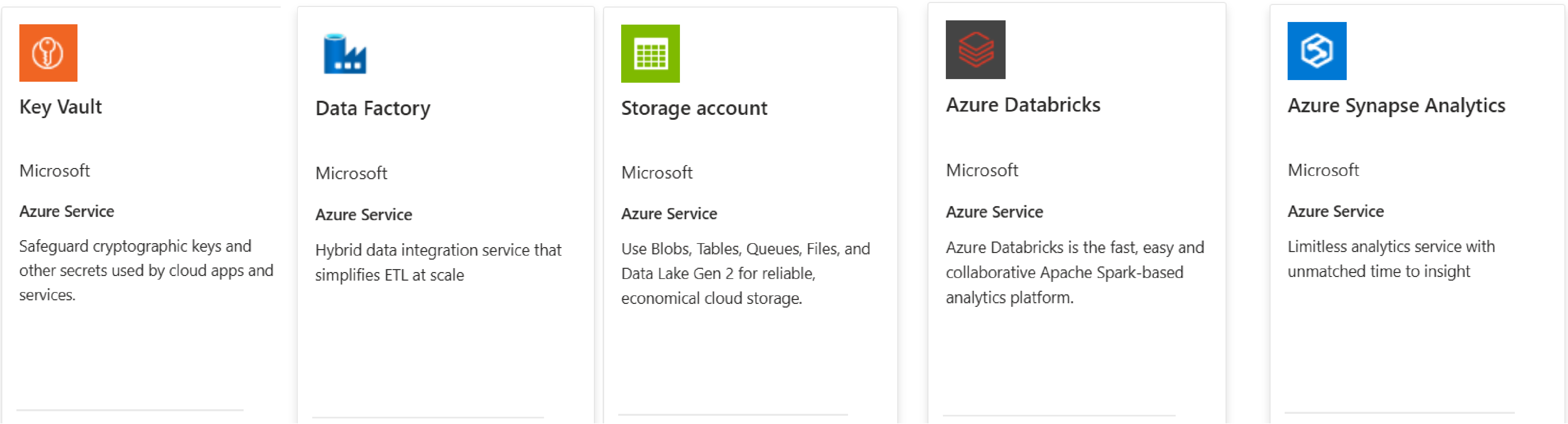
Analytics

Why Cloud is required for Big Data Solution

➢Cloud computing offers an effective solution towards dealing with big size information sets. Organizations can store their big-data efficiently manage them as well analyze them by leveraging scalability provided through clouds on demand resources such as storage capacity .

1. **Scalability**
2. **Cost Effectiveness**
3. **Performance**
4. **Accessibility**
5. **Security**

Azure Business Analytics Service Market Place



Key Vault

Microsoft

Azure Service

Safeguard cryptographic keys and other secrets used by cloud apps and services.

Data Factory

Microsoft

Azure Service

Hybrid data integration service that simplifies ETL at scale

Storage account

Microsoft

Azure Service

Use Blobs, Tables, Queues, Files, and Data Lake Gen 2 for reliable, economical cloud storage.

Azure Databricks

Microsoft

Azure Service

Azure Databricks is the fast, easy and collaborative Apache Spark-based analytics platform.

Azure Synapse Analytics

Microsoft

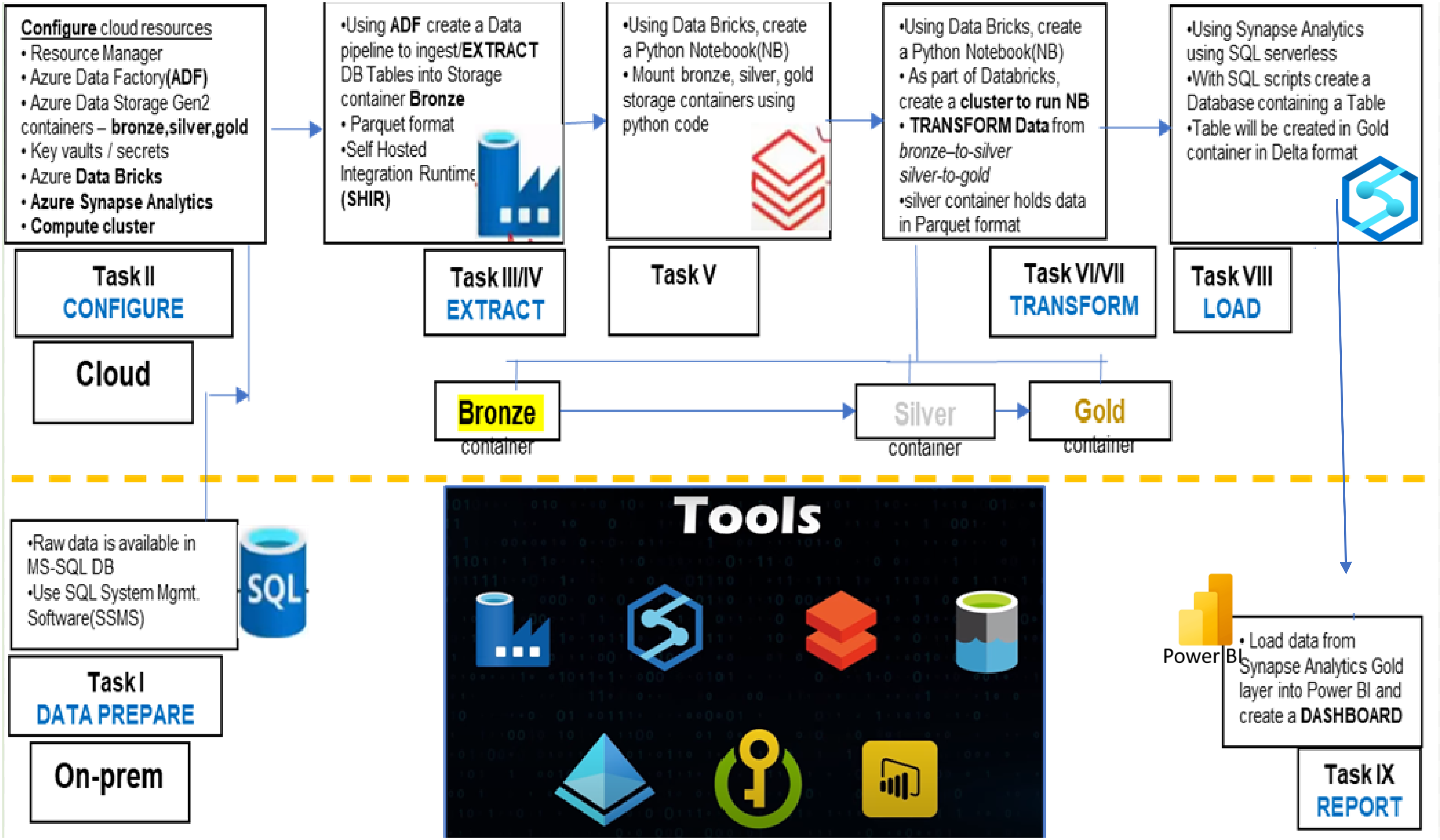
Azure Service

Limitless analytics service with unmatched time to insight

Data Engineering Solution Description

1. Given a big industrial dataset spread across different Storage repository in different types and different formats, build a Big data lake on Azure cloud.
2. Using the Azure data services and using the Extract, Transform, Load process, convert the raw data into a clean data and extract insights from the data and present the valuable data to the senior management for making business decisions.

Azure Data Engineering End to End Solution Architecture



Configure cloud resources
- Resource Manager
- Azure Data Factory(**ADF**)
- Azure Data Storage Gen2 containers – **bronze,silver,gold**
- Key vaults / secrets
- Azure **Data Bricks**
- **Azure Synapse Analytics**
- **Compute cluster**

Task II
CONFIGURE

Cloud

- Using **ADF** create a Data pipeline to ingest/**EXTRACT** DB Tables into Storage container **Bronze**
- Parquet format
- Self Hosted Integration Runtime (**SHIR**)

Task III/IV
EXTRACT

- Using Data Bricks, create a Python Notebook(NB)
- Mount bronze, silver, gold storage containers using python code

Task V

- Using Data Bricks, create a Python Notebook(NB)
- As part of Databricks, create a **cluster to run NB**
- **TRANSFORM Data** from *bronze–to-silver* *silver-to-gold*
- silver container holds data in Parquet format

Task VI/VII
TRANSFORM

- Using Synapse Analytics using SQL serverless
- With SQL scripts create a Database containing a Table
- Table will be created in Gold container in Delta format

Task VIII
LOAD

Bronze container

Silver container

Gold container

- Raw data is available in MS-SQL DB
- Use SQL System Mgmt. Software(SSMS)

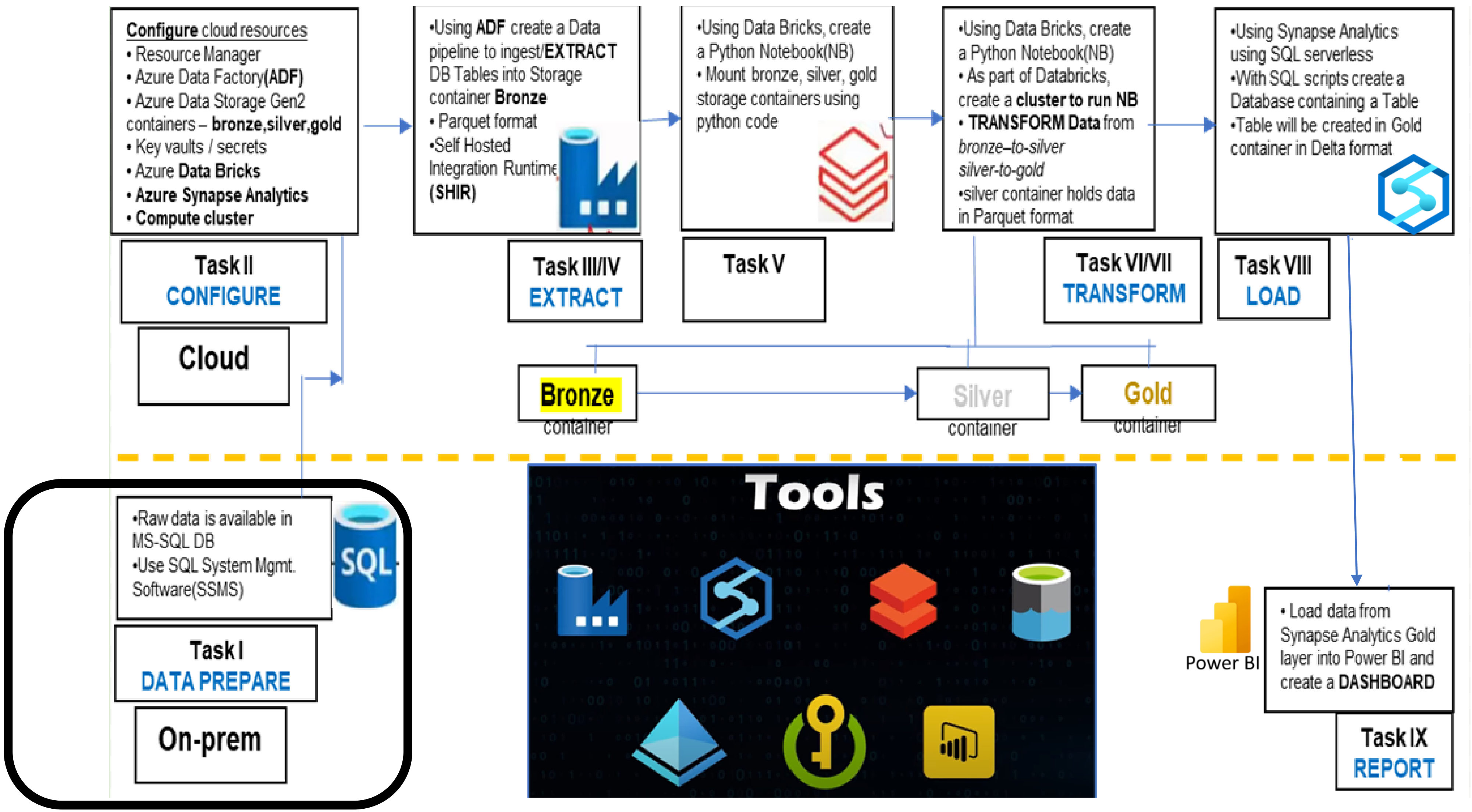
Task I
DATA PREPARE

On-prem

- Load data from Synapse Analytics Gold layer into Power BI and create a **DASHBOARD**

Task IX
REPORT

Azure Data Engineering End to End Project Architecture



Configure cloud resources
- Resource Manager
- Azure Data Factory(**ADF**)
- Azure Data Storage Gen2 containers – **bronze,silver,gold**
- Key vaults / secrets
- Azure **Data Bricks**
- **Azure Synapse Analytics**
- **Compute cluster**

Task II
CONFIGURE

Cloud

- Using **ADF** create a Data pipeline to ingest/**EXTRACT** DB Tables into Storage container **Bronze**
- Parquet format
- Self Hosted Integration Runtime (**SHIR**)

Task III/IV
EXTRACT

- Using Data Bricks, create a Python Notebook(NB)
- Mount bronze, silver, gold storage containers using python code

Task V

- Using Data Bricks, create a Python Notebook(NB)
- As part of Databricks, create a **cluster to run NB**
- **TRANSFORM Data** from *bronze–to-silver* *silver-to-gold*
- silver container holds data in Parquet format

Task VI/VII
TRANSFORM

- Using Synapse Analytics using SQL serverless
- With SQL scripts create a Database containing a Table
- Table will be created in Gold container in Delta format

Task VIII
LOAD

Bronze container

Silver container

Gold container

- Raw data is available in MS-SQL DB
- Use SQL System Mgmt. Software(SSMS)

SQL

Task I
DATA PREPARE

On-prem

Tools

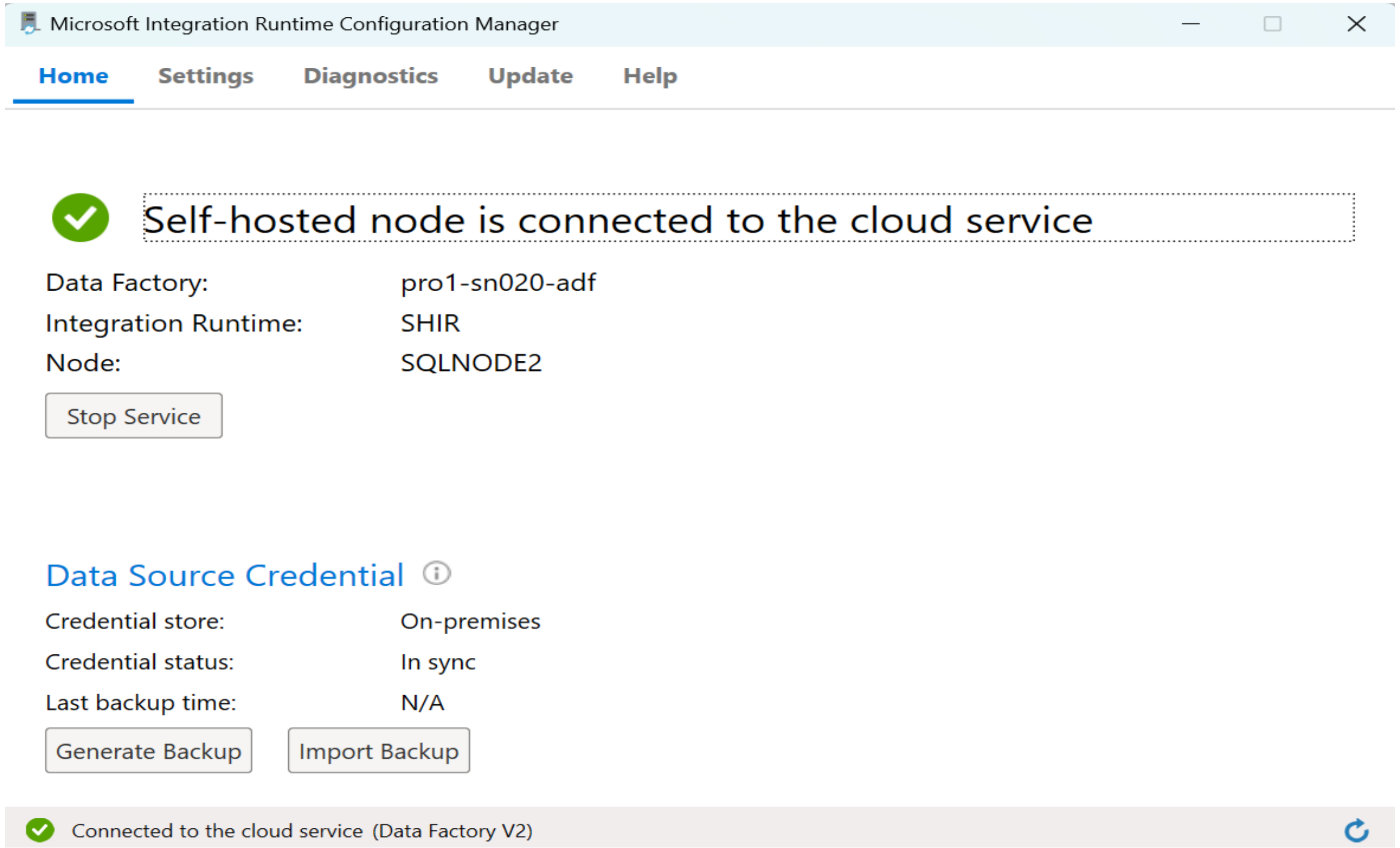
Power BI

- Load data from Synapse Analytics Gold layer into Power BI and create a **DASHBOARD**

Task IX
REPORT

On-Prem and Azure Cloud Resources

1. **OnPrem:** Raw Data available in **MS-SQL-DB**. Use **Self Hosted Integration Run Time** Migrated Data from On Prem to Microsoft Azure Cloud .
2. **Data Migration**



Microsoft Integration Runtime Configuration Manager

Self-hosted node is connected to the cloud service

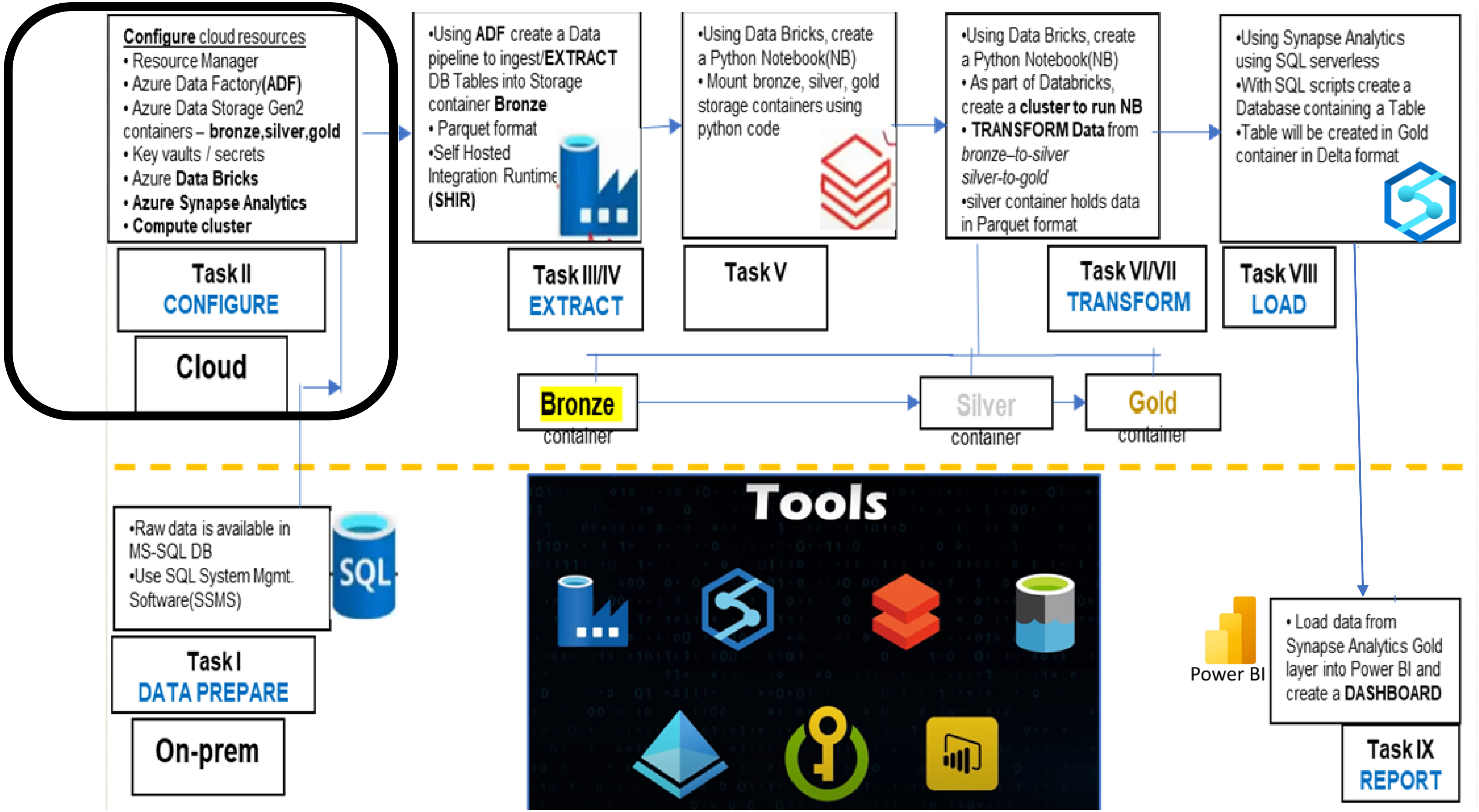
Data Factory: pro1-sn020-adf
Integration Runtime: SHIR
Node: SQLNODE2

Data Source Credential

Credential store: On-premises
Credential status: In sync
Last backup time: N/A

Connected to the cloud service (Data Factory V2)

Azure Data Engineering End to End Project Architecture



Configure cloud resources
- Resource Manager
- Azure Data Factory**(ADF)**
- Azure Data Storage Gen2 containers – **bronze,silver,gold**
- Key vaults / secrets
- Azure **Data Bricks**
- **Azure Synapse Analytics**
- **Compute cluster**

Task II
CONFIGURE

Cloud

- Using **ADF** create a Data pipeline to ingest/**EXTRACT** DB Tables into Storage container **Bronze**
- Parquet format
- Self Hosted Integration Runtime (**SHIR**)

Task III/IV
EXTRACT

- Using Data Bricks, create a Python Notebook(NB)
- Mount bronze, silver, gold storage containers using python code

Task V

- Using Data Bricks, create a Python Notebook(NB)
- As part of Databricks, create a **cluster to run NB**
- **TRANSFORM Data** from *bronze–to-silver* *silver-to-gold*
- silver container holds data in Parquet format

Task VI/VII
TRANSFORM

- Using Synapse Analytics using SQL serverless
- With SQL scripts create a Database containing a Table
- Table will be created in Gold container in Delta format

Task VIII
LOAD

Bronze container

Silver container

Gold container

Tools

- Raw data is available in MS-SQL DB
- Use SQL System Mgmt. Software(SSMS)

SQL

Task I
DATA PREPARE

On-prem

Power BI

- Load data from Synapse Analytics Gold layer into Power BI and create a **DASHBOARD**

Task IX
REPORT

On-Prem and Azure Cloud Resources

1. Azure Cloud Resources:

- Resource Manager
  - Key Vault
    - Secrets
  - Data Factory
    - Self Host Integration Run Time
  - Storage Account
    - Container
  - Data Bricks
    - Cluster
  - Synapse Workspace
    - SQL DB

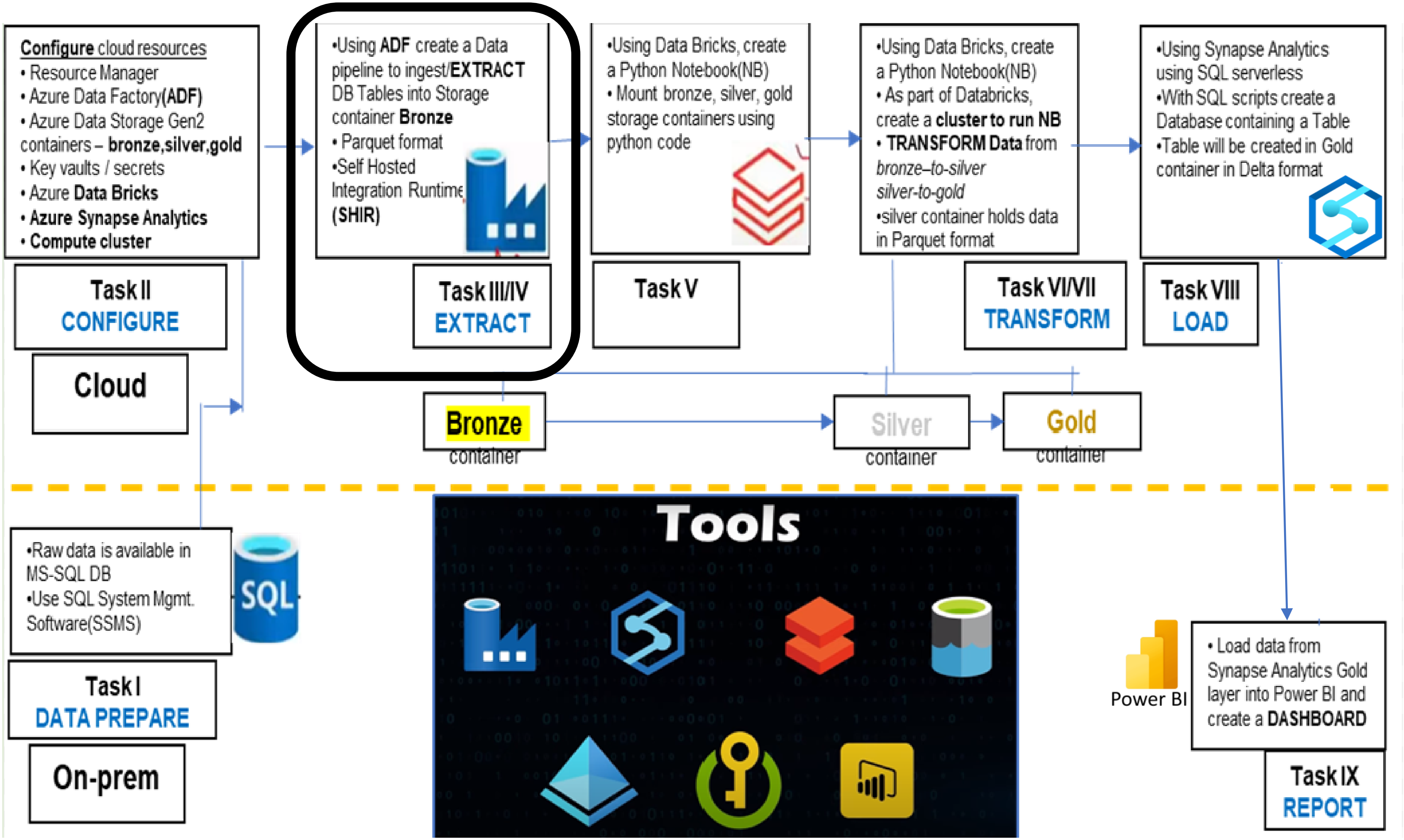
Resource Group

1. A **resource group is a container** that holds related resources for an Azure solution.
2. The resource group can include all the resources for the solution
3. Generally, add resources that share the same lifecycle to the same resource group so you can easily **deploy, update, and delete** them as a group
4. The resource group **stores metadata** about the resources.
5. Therefore, when you specify a location for the resource group, you're specifying where that **metadata is stored**.
6. you need to ensure that your data is stored in a particular region. Note that resources inside a resource group can be of different regions.

Azure S/W Building Blocks - Key Vault

- **Azure Key Vault** is a cloud service provided by Microsoft Azure to securely store and access secrets, such as API keys, passwords, certificates, and other sensitive data.
  1. **Secure Storage:** Encrypts secrets and keys with hardware security modules (HSMs).
  2. **Access Control:** Integration with Azure Active Directory (Azure AD) ensures fine-grained permissions to access secrets.

Azure Data Engineering End to End Project Architecture



Configure cloud resources
- Resource Manager
- Azure Data Factory(**ADF**)
- Azure Data Storage Gen2 containers – **bronze,silver,gold**
- Key vaults / secrets
- Azure **Data Bricks**
- **Azure Synapse Analytics**
- **Compute cluster**

Task II
CONFIGURE

Cloud

- Using **ADF** create a Data pipeline to ingest/**EXTRACT** DB Tables into Storage container **Bronze**
- Parquet format
- Self Hosted Integration Runtime (**SHIR**)

Task III/IV
EXTRACT

- Using Data Bricks, create a Python Notebook(NB)
- Mount bronze, silver, gold storage containers using python code

Task V

- Using Data Bricks, create a Python Notebook(NB)
- As part of Databricks, create a **cluster to run NB**
- **TRANSFORM Data** from *bronze–to-silver* *silver-to-gold*
- silver container holds data in Parquet format

Task VI/VII
TRANSFORM

- Using Synapse Analytics using SQL serverless
- With SQL scripts create a Database containing a Table
- Table will be created in Gold container in Delta format

Task VIII
LOAD

Bronze container

Silver container

Gold container

- Raw data is available in MS-SQL DB
- Use SQL System Mgmt. Software(SSMS)

Task I
DATA PREPARE

On-prem

Tools

- Load data from Synapse Analytics Gold layer into Power BI and create a **DASHBOARD**

Task IX
REPORT

Azure S/W Building Blocks – Data Factory(ADF)

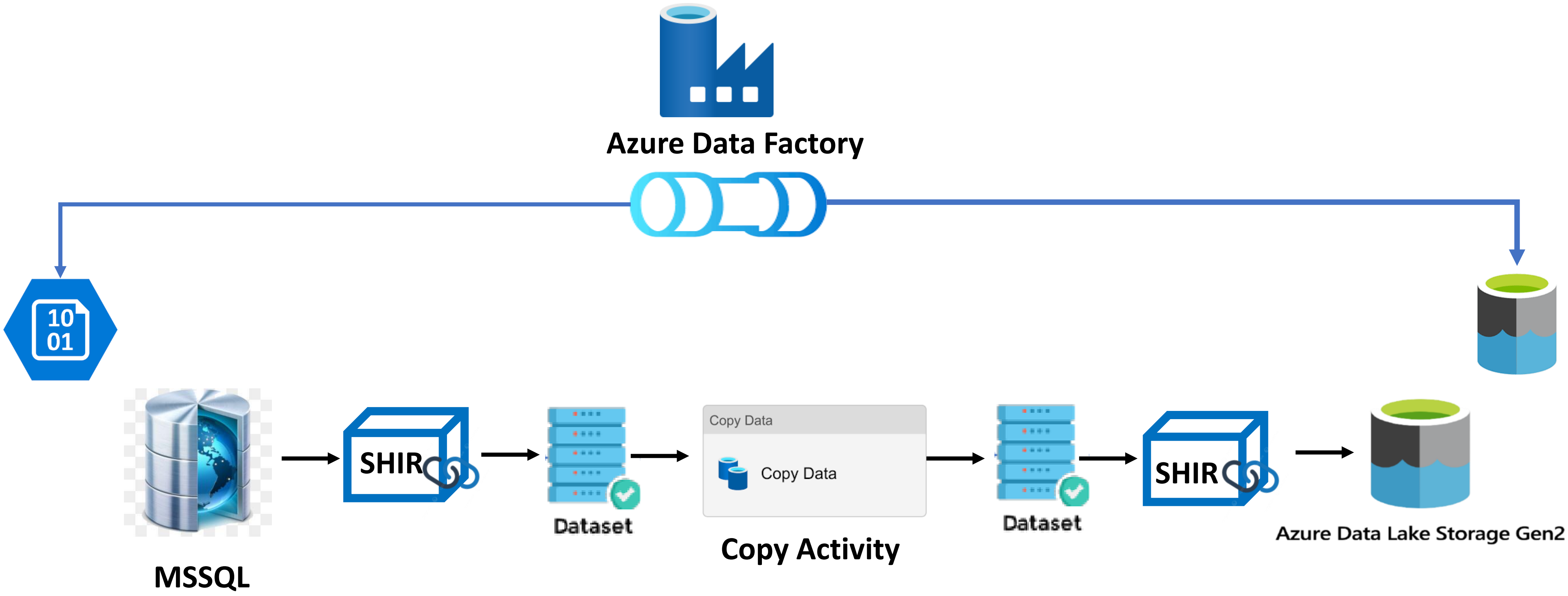
1. Data Factory is a cloud-based **ETL (Extract, Transform, Load)** service that enables the movement, transformation, and loading of data from various sources to different destinations.
2. Loosely, we can imagine Data Factory as a pipeline editor.

Types of Integration Runtime in Data Factory (ADF)

1. **Azure IR:** For data movement/transformation within Azure.
2. **Self-Hosted IR:** For on-premises data integration.
3. **Azure-SSIS IR:** To run SSIS packages in the cloud.
    - In our projects so far we have used only 1 and 2. We have not used Azure SSIS IR.

Connectivity Pipeline example

data migration from on Prem MSSQL to Azure Cloud Storage



Components of Azure Data Factory - I

1. **Datasets:** Structures representing data to be used in activities.
2. **linked service** defines a **connection string** or parameters needed to connect to a data source.
3. **Activities:** Tasks performed inside pipelines.
4. **Integration Runtime** in Azure is a compute infrastructure used by Azure Data Factory to **perform data movement, transformation**, and integration tasks across different network environments.

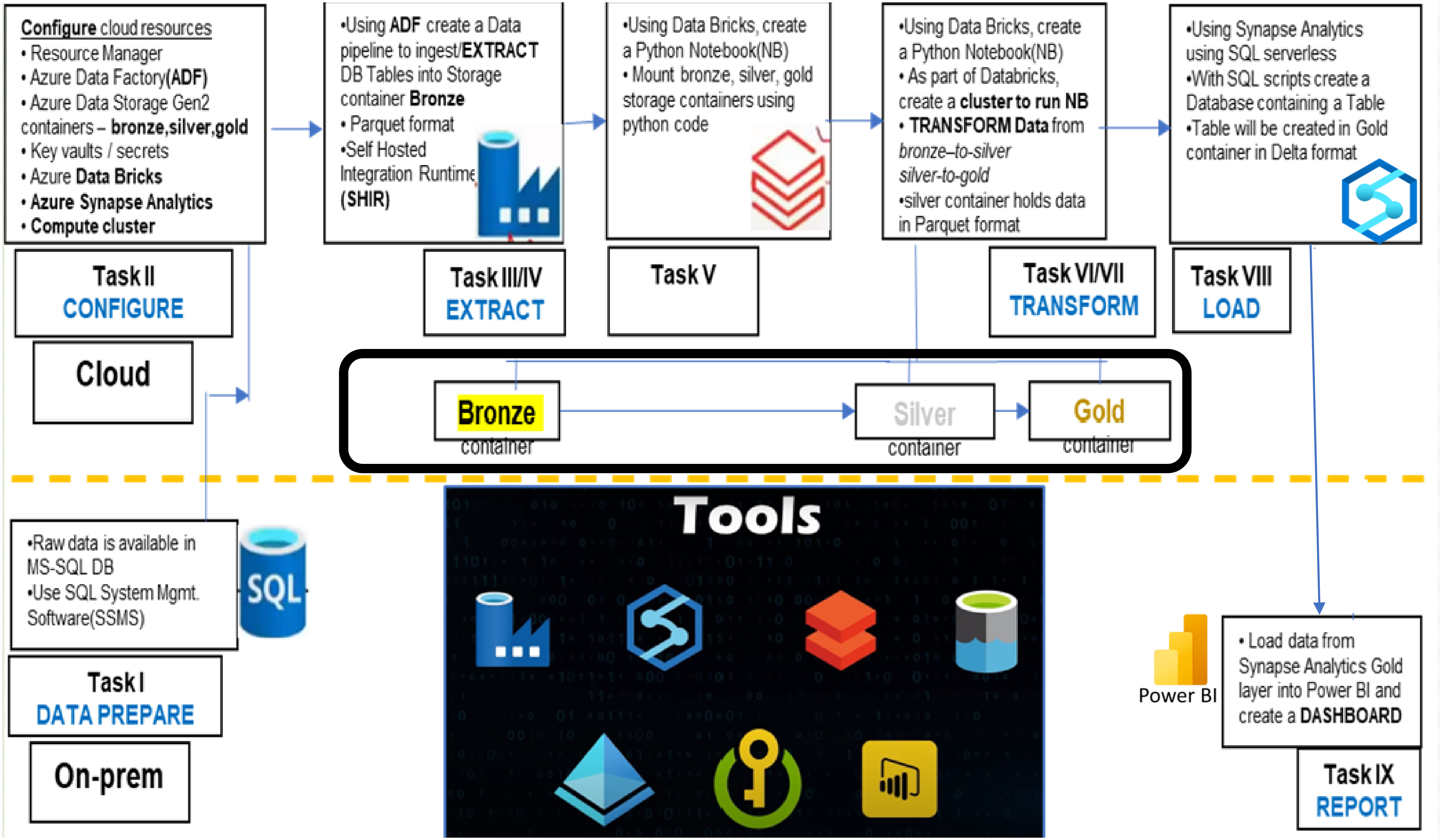
Components of Azure Data Factory - II

5. **Copy Data** is a tool that moves data from one place to another in Azure.
6. **Triggers** are used to **schedule pipeline execution**, e.g., **Scheduled Triggers**, **Event-Based Triggers**.
7. **Debugging** is the process of testing and validating pipelines before publishing.

Components of Azure Data Factory - III

8. **For Each** is a loop activity that performs actions on each item in a list
9. **Lookup** is an activity in Azure Data Factory that retrieves data from a dataset and passes it to subsequent steps in a pipeline.
10. **SQL Server** is a fully managed relational database service in Azure that allows you to store, manage, and query data using SQL.

Azure Data Engineering End to End Project Architecture



<u>Configure</u> cloud resources

- Resource Manager
- Azure Data Factory(**ADF**)
- Azure Data Storage Gen2 containers – **bronze,silver,gold**
- Key vaults / secrets
- Azure **Data Bricks**
- **Azure Synapse Analytics**
- **Compute cluster**

Task II
CONFIGURE

Cloud

- Using **ADF** create a Data pipeline to ingest/**EXTRACT** DB Tables into Storage container **Bronze**
- Parquet format
- Self Hosted Integration Runtime (**SHIR**)

Task III/IV
EXTRACT

- Using Data Bricks, create a Python Notebook(NB)
- Mount bronze, silver, gold storage containers using python code

Task V

- Using Data Bricks, create a Python Notebook(NB)
- As part of Databricks, create a **cluster to run NB**
- **TRANSFORM Data** from *bronze–to-silver* *silver-to-gold*
- silver container holds data in Parquet format

Task VI/VII
TRANSFORM

- Using Synapse Analytics using SQL serverless
- With SQL scripts create a Database containing a Table
- Table will be created in Gold container in Delta format

Task VIII
LOAD

Bronze container

Silver container

Gold container

- Raw data is available in MS-SQL DB
- Use SQL System Mgmt. Software(SSMS)

Task I
DATA PREPARE

On-prem

Power BI

- Load data from Synapse Analytics Gold layer into Power BI and create a **DASHBOARD**

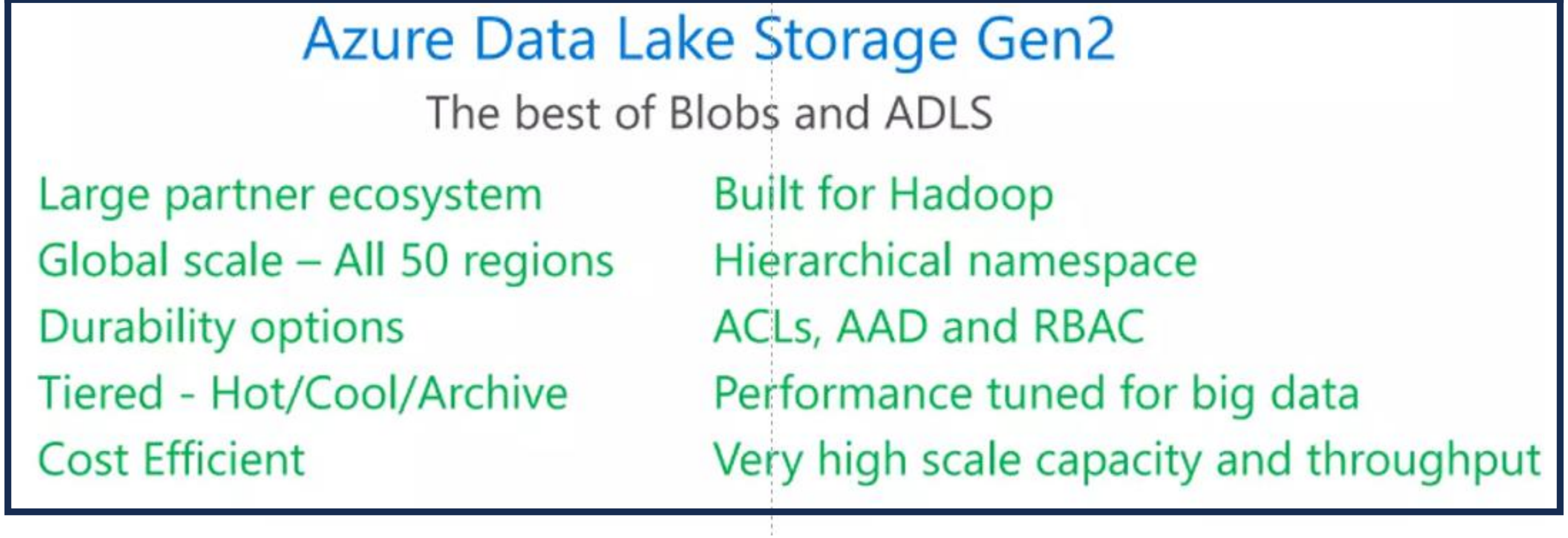
Task IX
REPORT

Azure S/W Building Blocks Part 3 – Storage Account

1. **Azure Storage** is a cloud service that provides scalable and secure storage for data, while a container in Azure Storage is a logical unit used to organize and store blob data, such as files or objects.

Azure S/W Building Blocks – Storage Account

Blob Storage General purpose object Storage	**Data Lake Store** Optimized from Big Data analytics



Azure Data Lake Storage Gen2

The best of Blobs and ADLS

- Large partner ecosystem
- Global scale – All 50 regions
- Durability options
- Tiered - Hot/Cool/Archive
- Cost Efficient

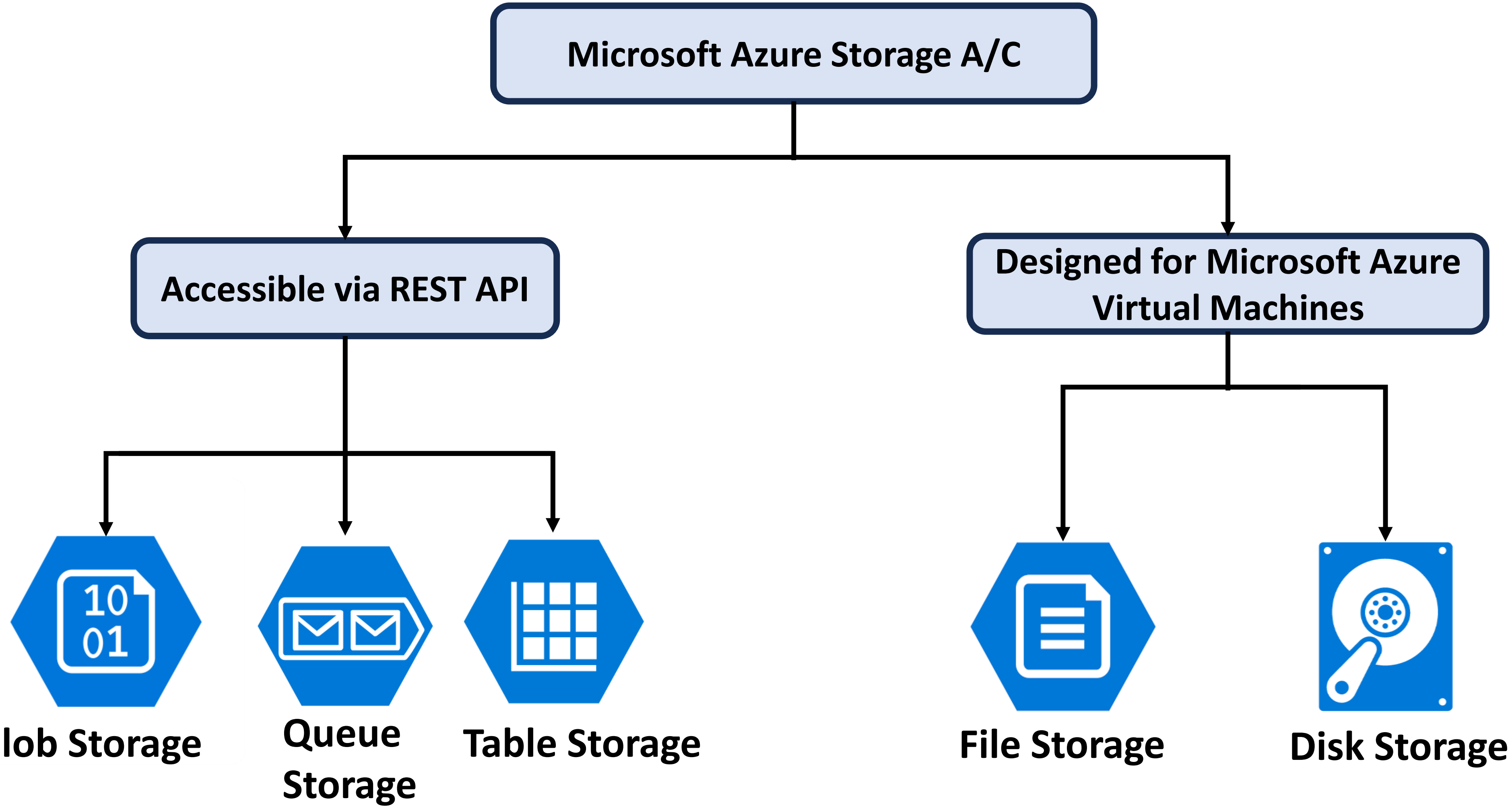
- Built for Hadoop
- Hierarchical namespace
- ACLs, AAD and RBAC
- Performance tuned for big data
- Very high scale capacity and throughput

Types of Storage

1. encryption scope for all **blobs in the container**
2. version-level immutability support

- **Azure Queue Storage** is a service for storing large numbers of messages. You access messages from anywhere in the world via authenticated calls using HTTP or HTTPS. A queue message can be up to 64 KB in size.
- **Table storage** offers higher performance and availability, global distribution, and automatic secondary indexes. It is also available in a consumption-based **serverless** mode
- **Azure Files** offers fully managed file shares in the cloud that are accessible via the Server Message Block (SMB) and Network File System (NFS) file system protocols.
- **Azure managed disks are block-level storage** volumes that are managed by Azure and used with Azure Virtual Machines. Managed disks are like physical disks in an on-premises server, but they're virtualized.

Types of Storage



Microsoft Azure Storage A/C

Accessible via REST API

Designed for Microsoft Azure Virtual Machines

Blob Storage

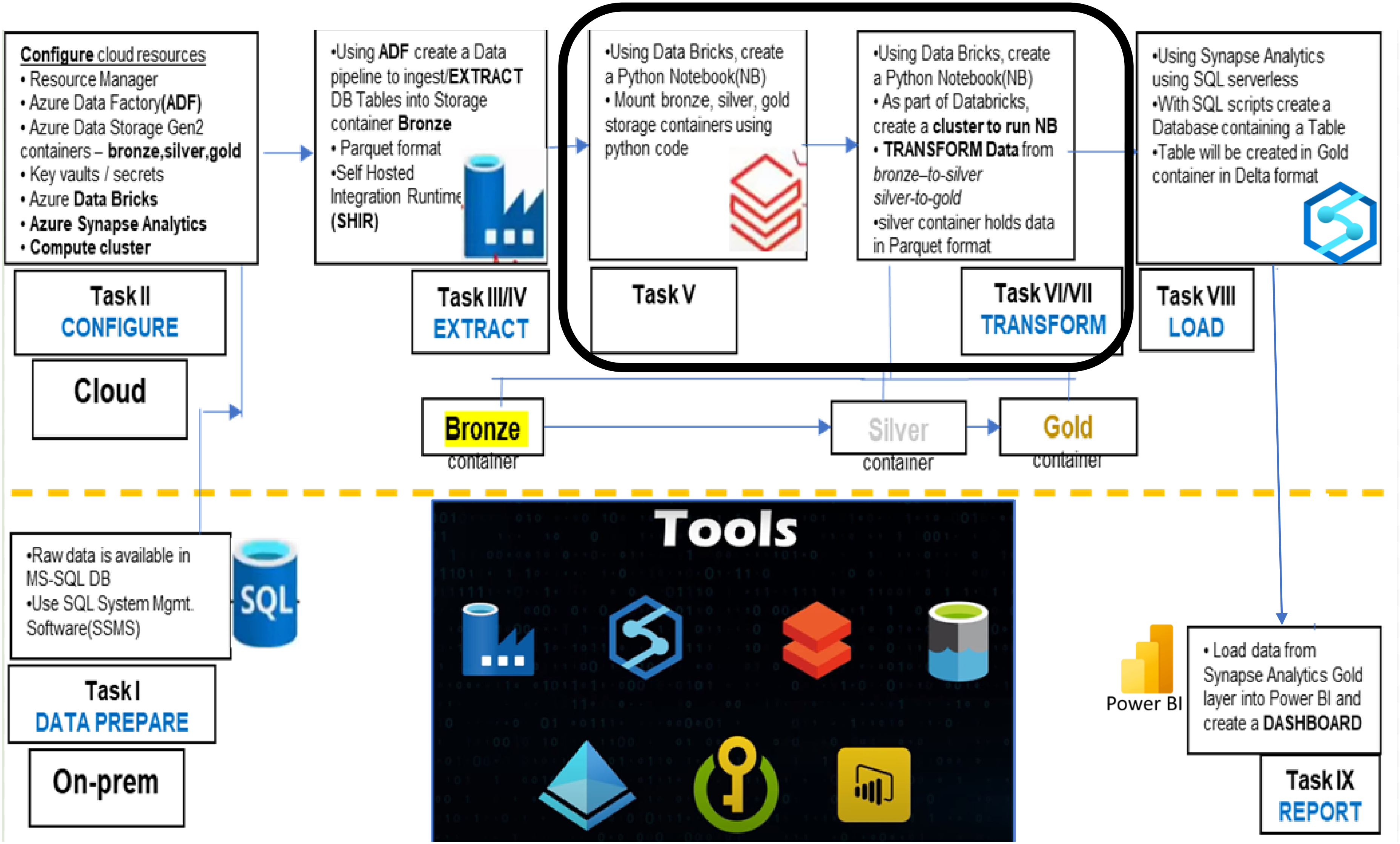
Queue Storage

Table Storage

File Storage

Disk Storage

Azure Data Engineering End to End Project Architecture



Configure cloud resources
- Resource Manager
- Azure Data Factory(ADF)
- Azure Data Storage Gen2 containers – bronze,silver,gold
- Key vaults / secrets
- Azure Data Bricks
- Azure Synapse Analytics
- Compute cluster

Task II
CONFIGURE

Cloud

- Using ADF create a Data pipeline to ingest/EXTRACT DB Tables into Storage container Bronze
- Parquet format
- Self Hosted Integration Runtime (SHIR)

Task III/IV
EXTRACT

- Using Data Bricks, create a Python Notebook(NB)
- Mount bronze, silver, gold storage containers using python code

Task V

- Using Data Bricks, create a Python Notebook(NB)
- As part of Databricks, create a cluster to run NB
- TRANSFORM Data from *bronze–to-silver* *silver-to-gold*
- silver container holds data in Parquet format

Task VI/VII
TRANSFORM

- Using Synapse Analytics using SQL serverless
- With SQL scripts create a Database containing a Table
- Table will be created in Gold container in Delta format

Task VIII
LOAD

Bronze container

Silver container

Gold container

- Raw data is available in MS-SQL DB
- Use SQL System Mgmt. Software(SSMS)

Task I
DATA PREPARE

On-prem

Tools

Power BI

- Load data from Synapse Analytics Gold layer into Power BI and create a DASHBOARD

Task IX
REPORT

Azure S/W Building Blocks – Azure Data Bricks

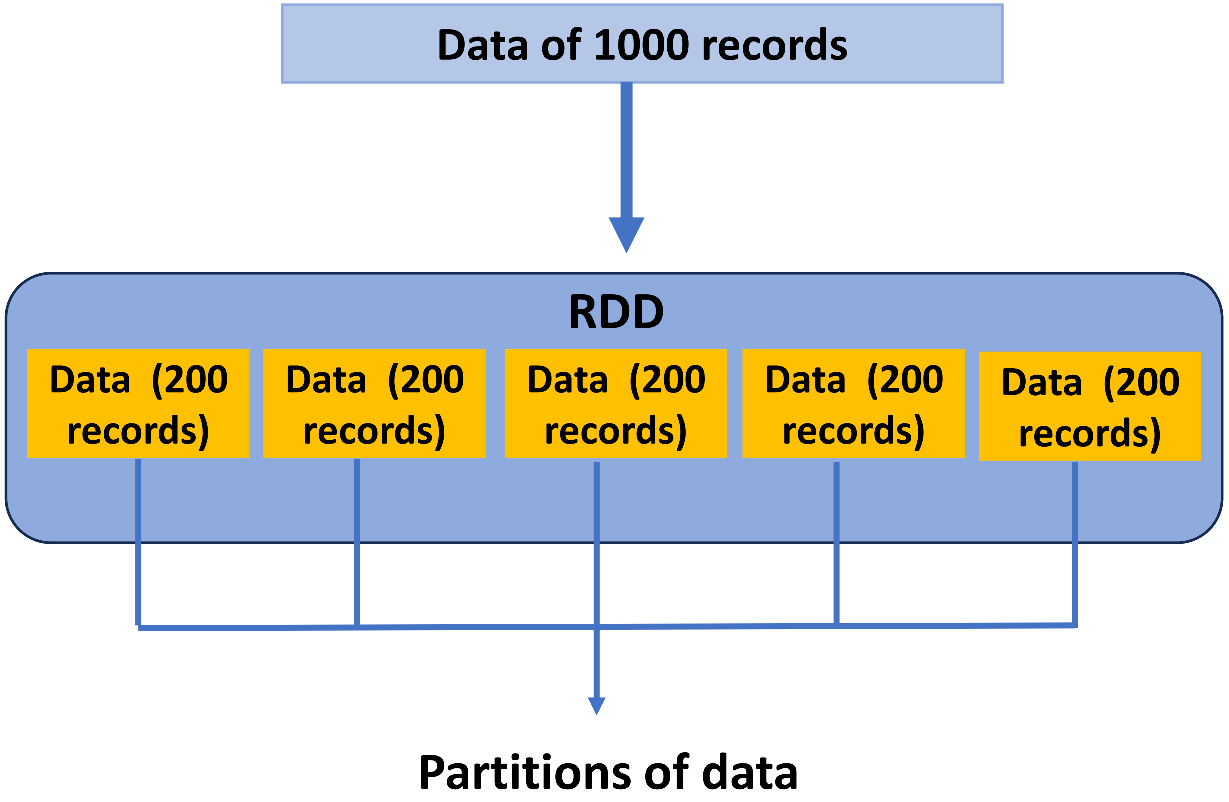
1. Databricks is a crucial analytics tool on Azure. It is offered on azure as software as a service (SaaS). It is powered by **Apache Spark.**
   - Data Bricks is also available outside azure as an independent software
2. The Software tool supports **Java, SQL, Scala, Python, R, and Scikit Learn**.
   - In our projects we use Python.

Apache Spark

1. Spark is the distributed computing engine behind Azure Databricks, which is used to **handle massive amounts of data** and carry out analytics.
2. By offering **high-level APIs and libraries** for diverse activities including **data transformation, machine learning, streaming analytics, and graph processing,** it makes it possible for data scientists, engineers, and analysts to work with enormous datasets effectively.
3. The essence, **Spark with Azure Databricks** makes it easier to process, explore, and do advanced analytics on data in a cloud environment that is scalable and collaborative.
4. Apache Spark Download Link
   - https://spark.apache.org/downloads.html
   - Spark Latest Version - Spark 3.4.4 released (Oct 27, 2024)

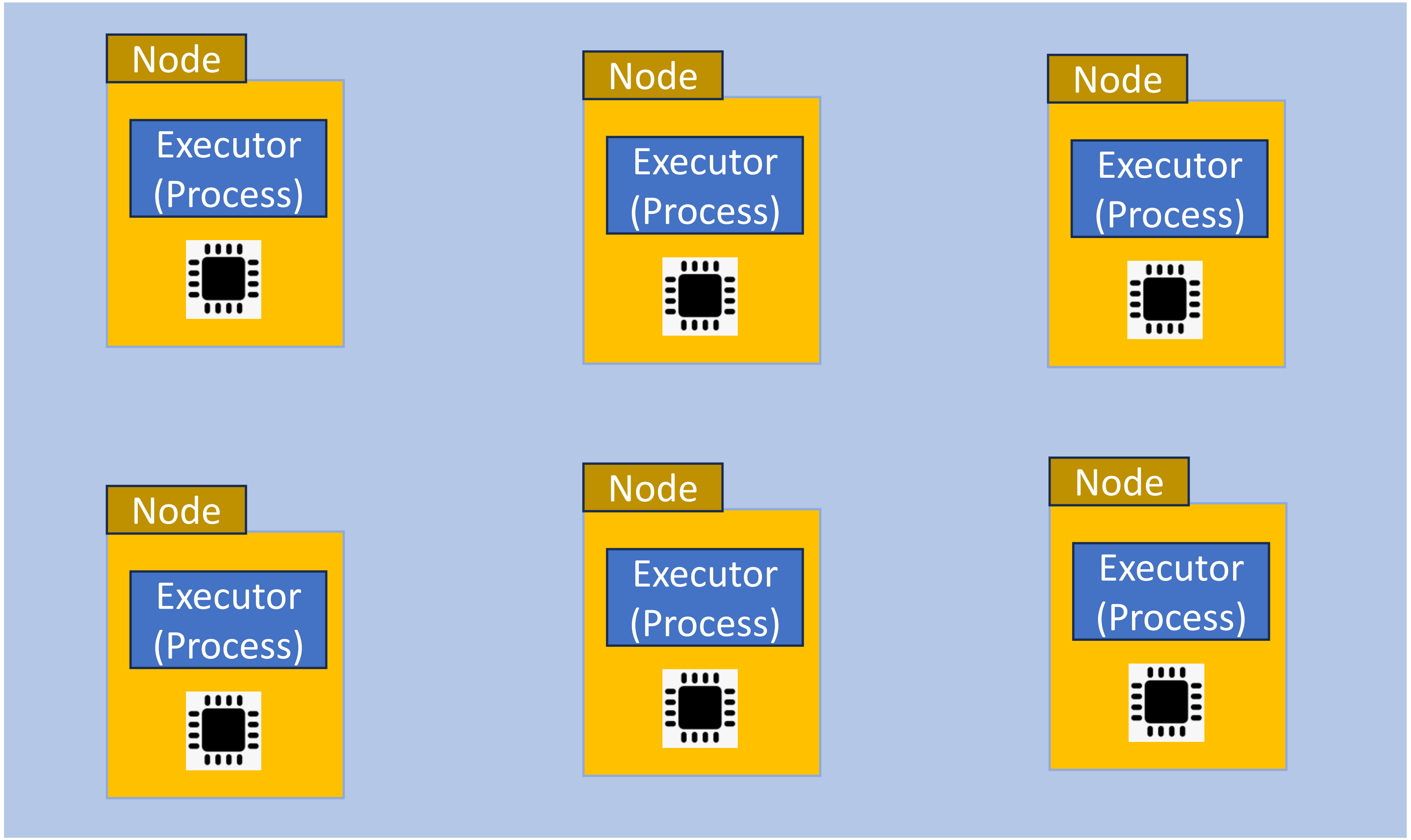
Spark Cluster

- RDD – Resilient Distributed Datasets
- The core abstraction spark provides is RDD

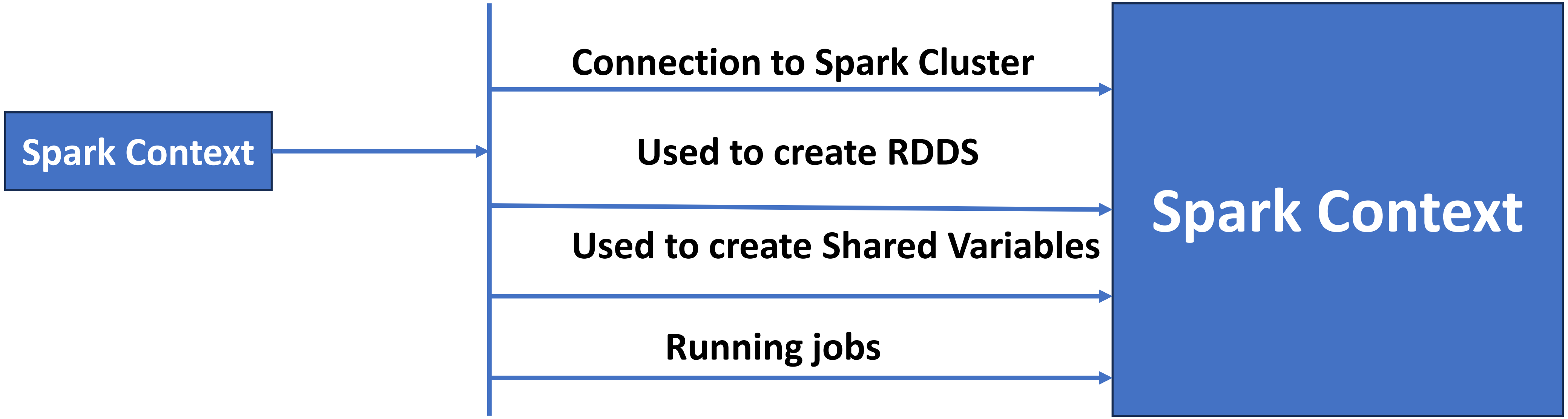


Sample : https://www.chaosgenius.io/blog/apache-spark-with-python/#resource-management

Spark Cluster



Spark Context



Spark Context

Connection to Spark Cluster

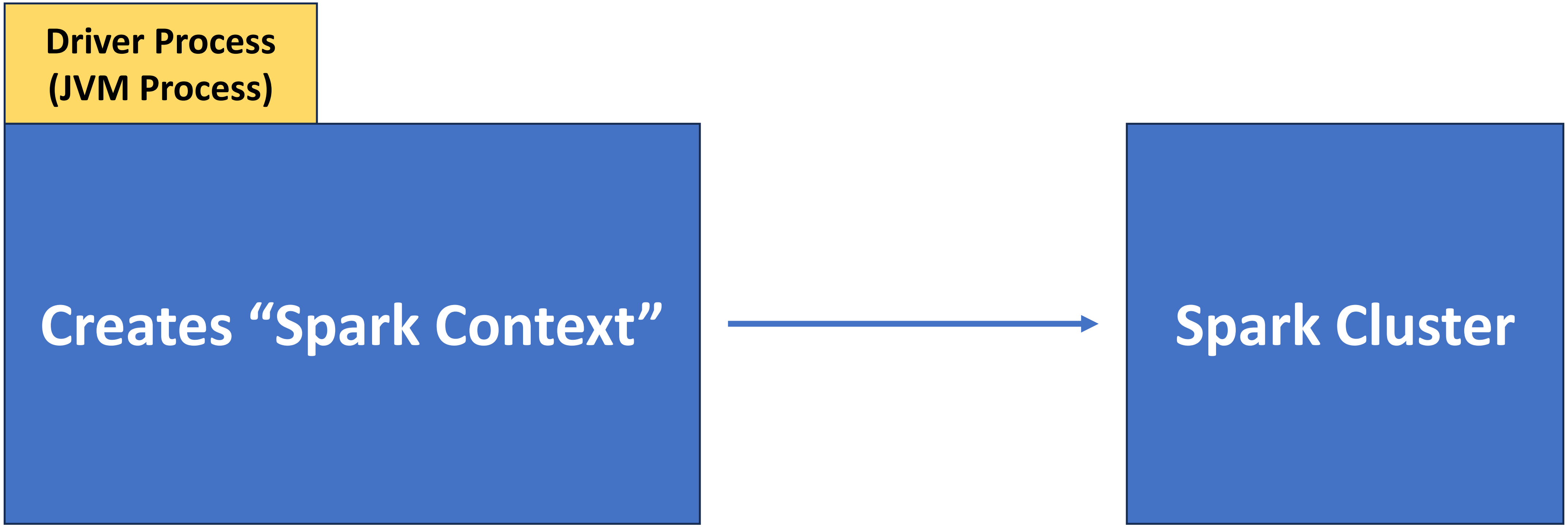
Used to create RDDS

Used to create Shared Variables

Running jobs

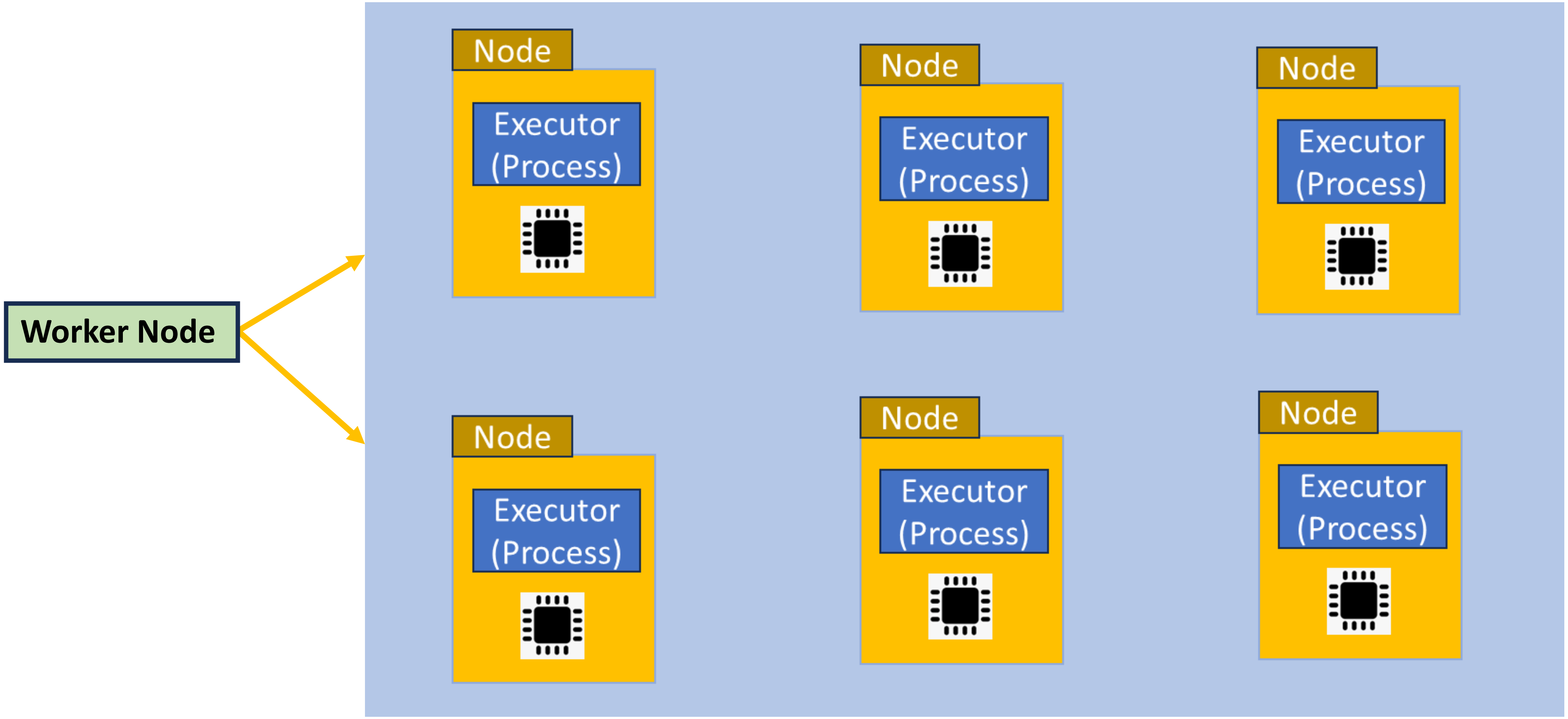
Spark Context

Driver Process



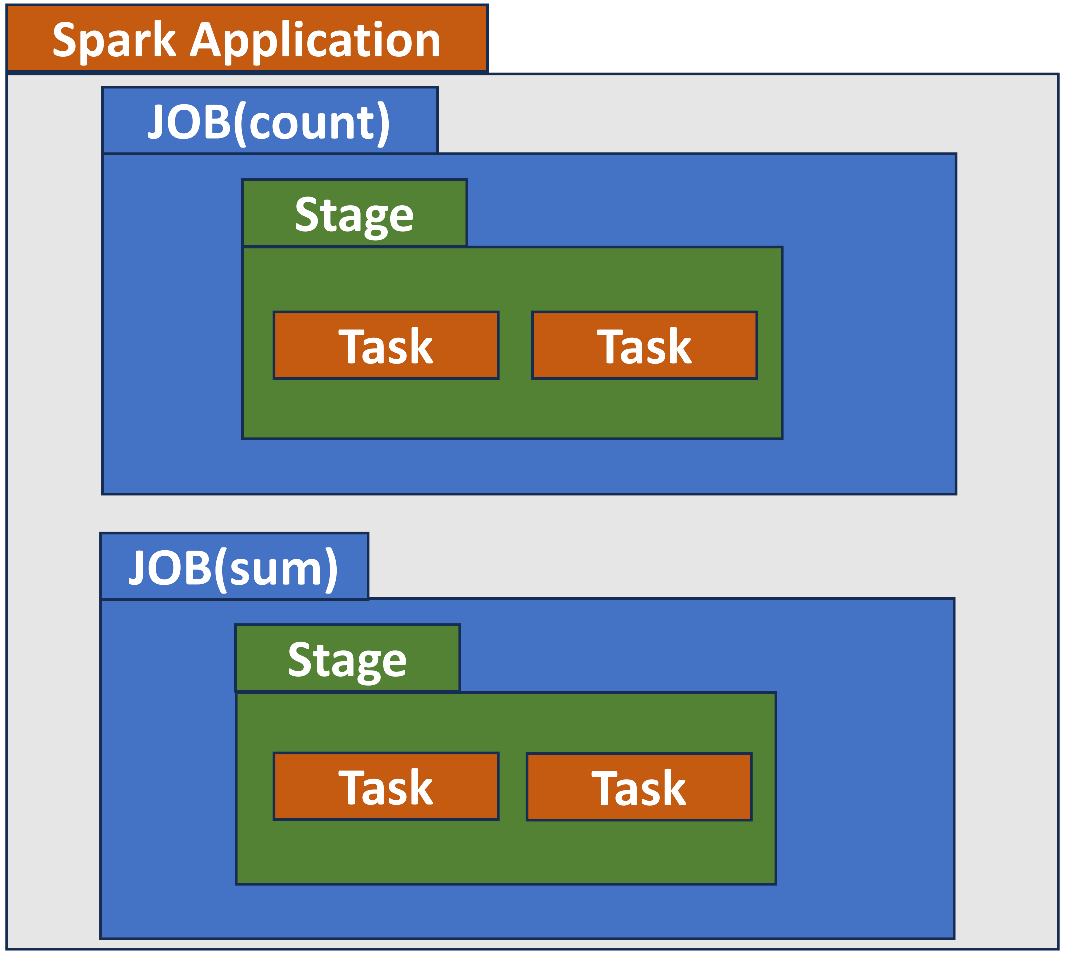
Executor

- Executor is a process Which runs on Worker Nodes

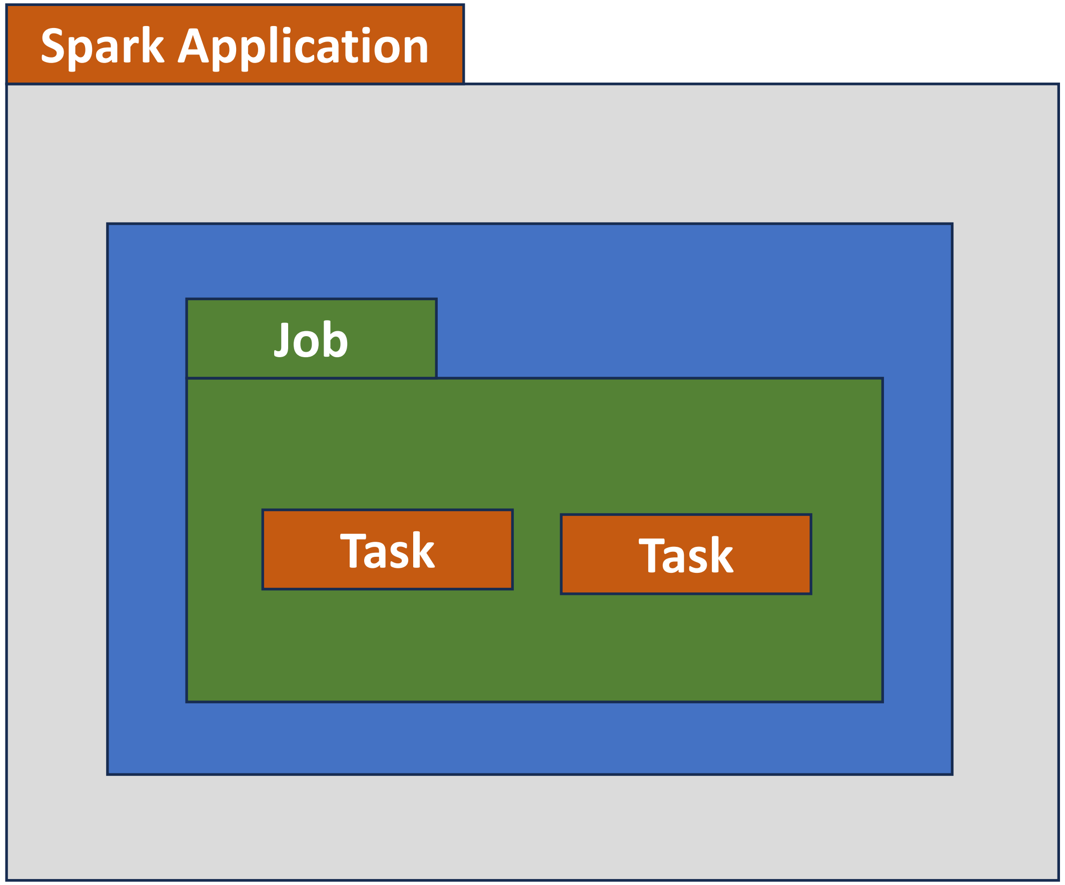


Stage and Task

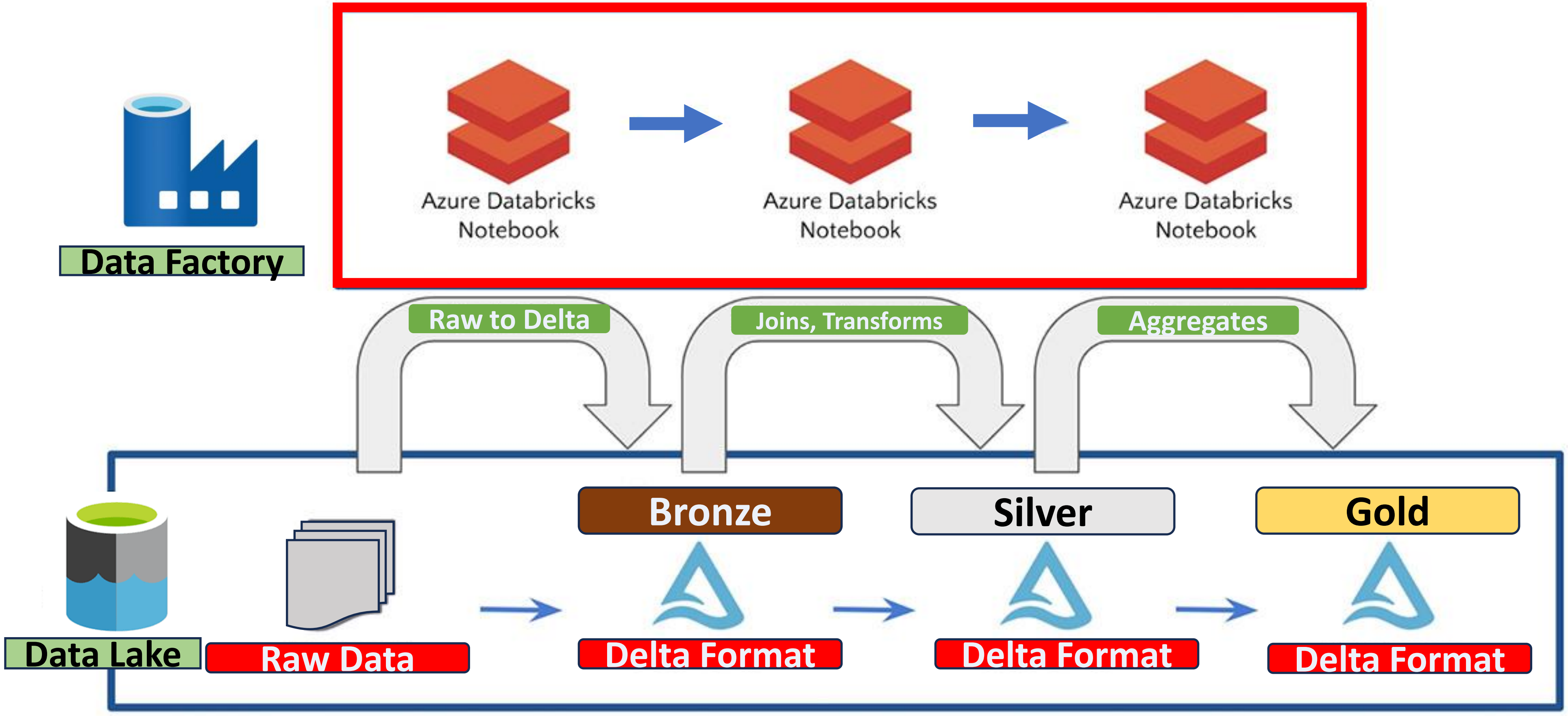
```
Val data = Array(1,2,3,4,5,6,7,8,9,10)
Val dotards = dotards(data, 2)
Val nonelements = dotards. Count
Val sum = dotards. Sum
```



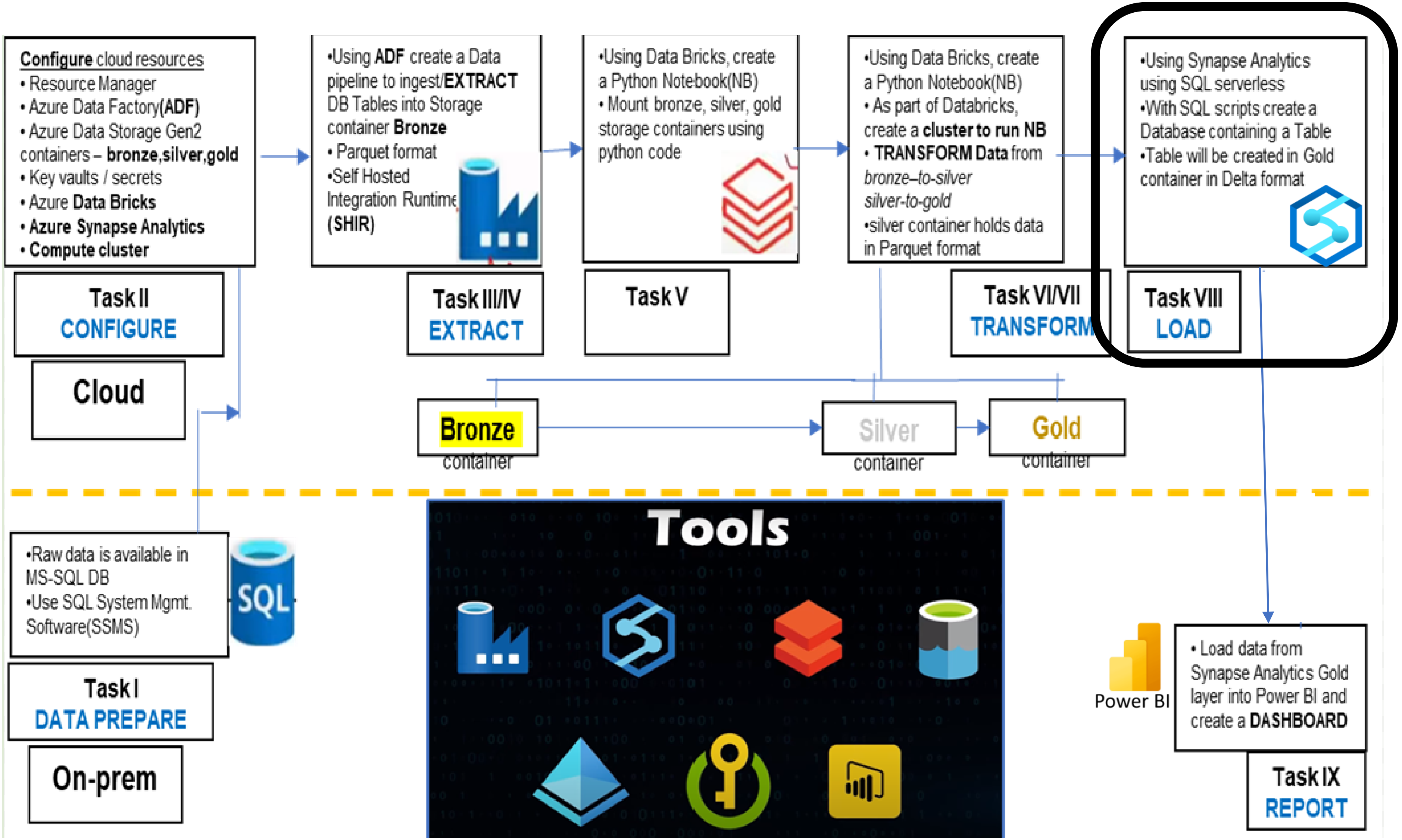
```
Val data = Array(1,2,3,4,5,6,7,8,9,10)
Val dotards = dotards(data, 2)
Val nonelements = dotards. Count
```



Data Bricks Data Transformations



Azure Data Engineering End to End Project Architecture



Configure cloud resources
- Resource Manager
- Azure Data Factory(**ADF**)
- Azure Data Storage Gen2 containers – **bronze,silver,gold**
- Key vaults / secrets
- Azure **Data Bricks**
- **Azure Synapse Analytics**
- **Compute cluster**

Task II
CONFIGURE

Cloud

- Using **ADF** create a Data pipeline to ingest/**EXTRACT** DB Tables into Storage container **Bronze**
- Parquet format
- Self Hosted Integration Runtime (**SHIR**)

Task III/IV
EXTRACT

- Using Data Bricks, create a Python Notebook(NB)
- Mount bronze, silver, gold storage containers using python code

Task V

- Using Data Bricks, create a Python Notebook(NB)
- As part of Databricks, create a **cluster to run NB**
- **TRANSFORM Data** from *bronze–to-silver* *silver-to-gold*
- silver container holds data in Parquet format

Task VI/VII
TRANSFORM

- Using Synapse Analytics using SQL serverless
- With SQL scripts create a Database containing a Table
- Table will be created in Gold container in Delta format

Task VIII
LOAD

Bronze container

Silver container

Gold container

- Raw data is available in MS-SQL DB
- Use SQL System Mgmt. Software(SSMS)

Task I
DATA PREPARE

On-prem

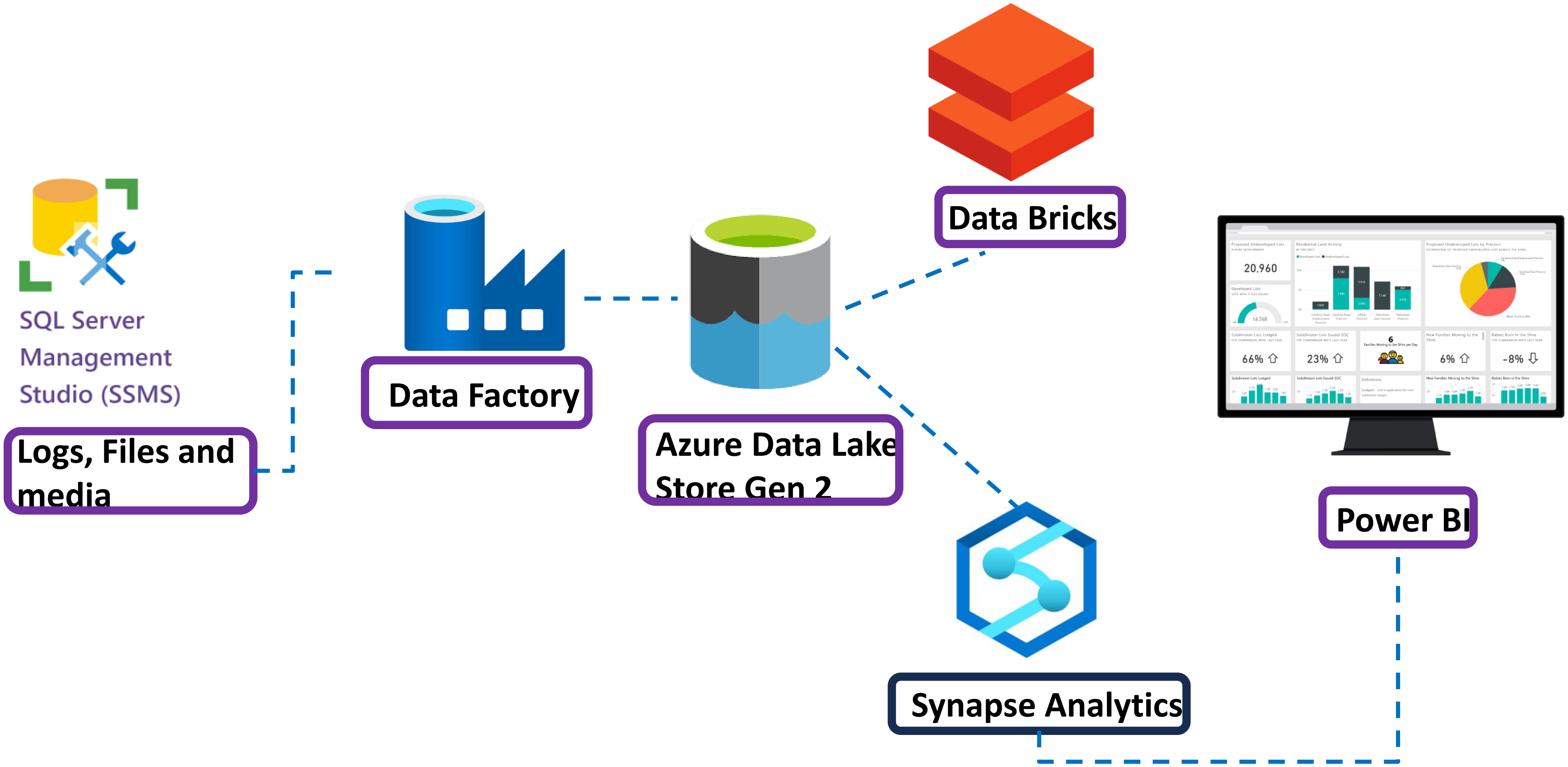
- Load data from Synapse Analytics Gold layer into Power BI and create a **DASHBOARD**

Task IX
REPORT

Azure S/W Building Blocks – Synapse Analytics

1. Synapse Workspace is an integrated **big data and data warehousing, SQL Database** in Synapse is a managed **relational database service**, and a stored procedure is a precompiled collection of SQL queries and logic that can be executed within the database.
2. Metadata refers to data that describes other data, such as the structure, properties, and relationships of data stored in Azure services.
3. Azure Synapse Analytics, **Synapse SQL** allows you to **create stored procedures**, which are a set of precompiled SQL queries that can be executed together. **Stored procedures in Synapse SQL** are particularly useful for encapsulating business logic, improving performance, and enabling reusable and modular SQL code.

Synapse Analytics Work Space Architecture



SQL Server Management Studio (SSMS)

Logs, Files and media

Data Factory

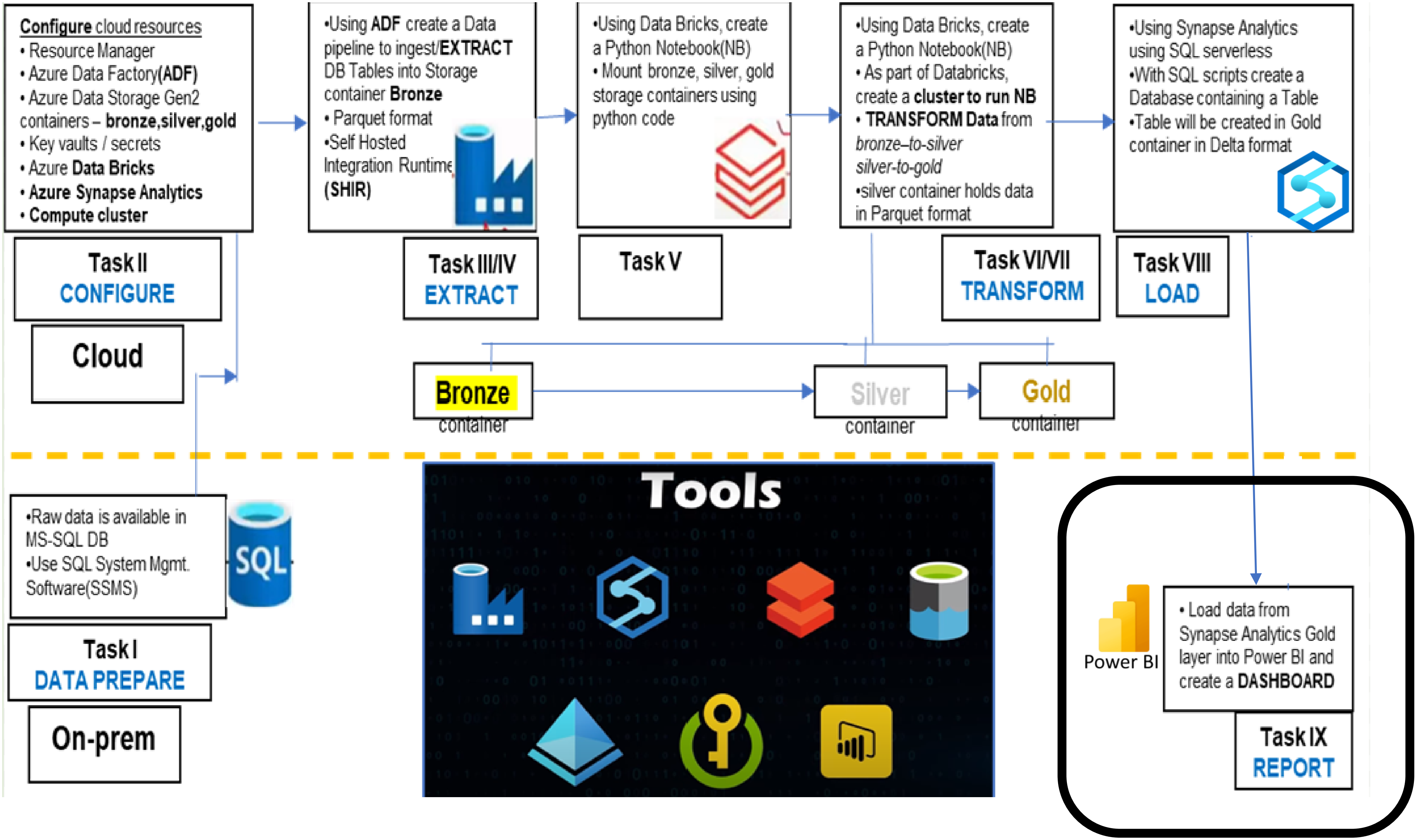
Azure Data Lake Store Gen 2

Data Bricks

Synapse Analytics

Power BI

Azure Data Engineering End to End Project Architecture



Configure cloud resources
- Resource Manager
- Azure Data Factory(ADF)
- Azure Data Storage Gen2 containers – bronze,silver,gold
- Key vaults / secrets
- Azure Data Bricks
- Azure Synapse Analytics
- Compute cluster

Task II
CONFIGURE

Cloud

- Using ADF create a Data pipeline to ingest/EXTRACT DB Tables into Storage container Bronze
- Parquet format
- Self Hosted Integration Runtime (SHIR)

Task III/IV
EXTRACT

- Using Data Bricks, create a Python Notebook(NB)
- Mount bronze, silver, gold storage containers using python code

Task V

- Using Data Bricks, create a Python Notebook(NB)
- As part of Databricks, create a cluster to run NB
- TRANSFORM Data from *bronze–to-silver* *silver-to-gold*
- silver container holds data in Parquet format

Task VI/VII
TRANSFORM

- Using Synapse Analytics using SQL serverless
- With SQL scripts create a Database containing a Table
- Table will be created in Gold container in Delta format

Task VIII
LOAD

Bronze container

Silver container

Gold container

- Raw data is available in MS-SQL DB
- Use SQL System Mgmt. Software(SSMS)

Task I
DATA PREPARE

On-prem

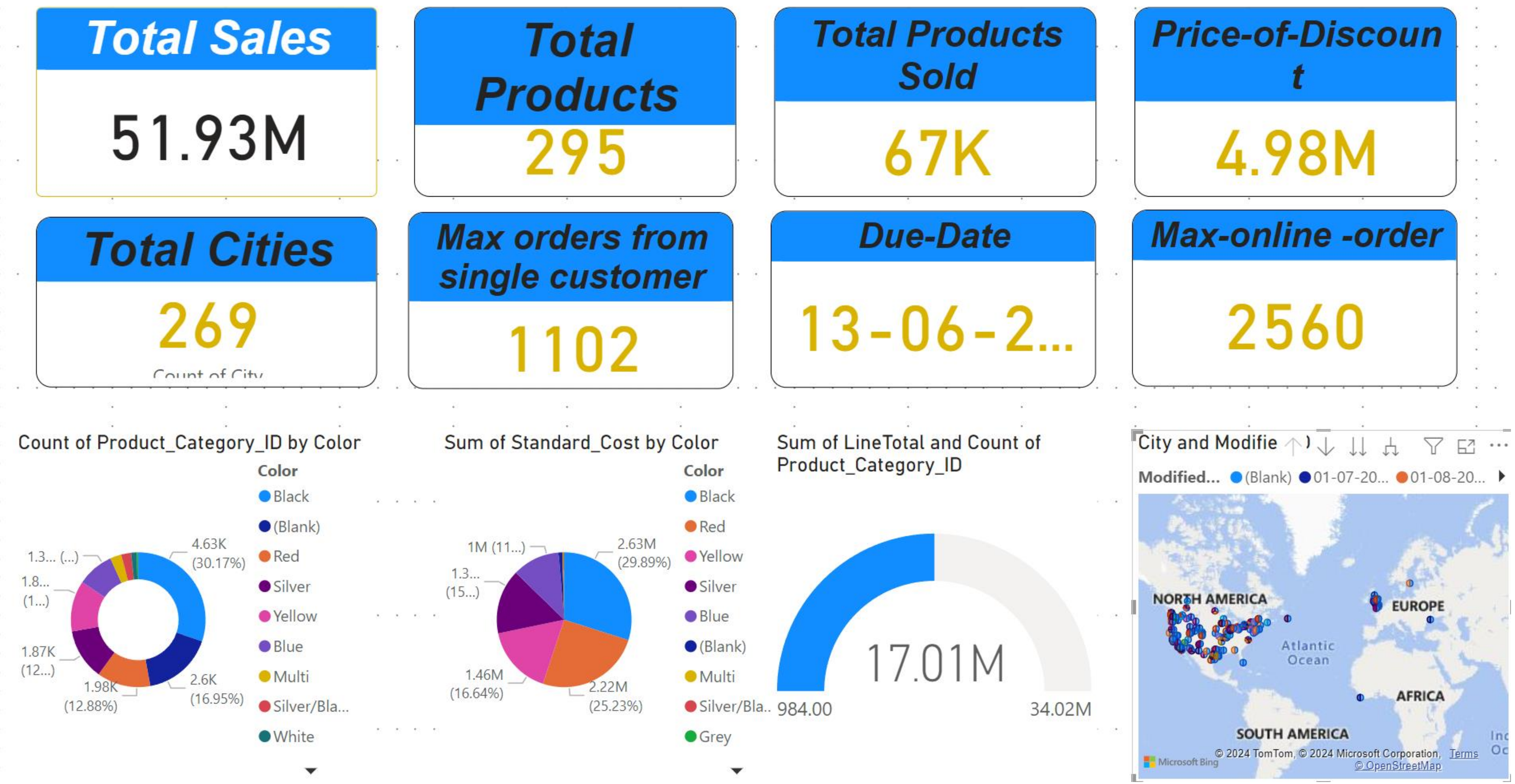
- Load data from Synapse Analytics Gold layer into Power BI and create a DASHBOARD

Task IX
REPORT

Power BI – OnPrem

1. Power BI is a **powerful business intelligence** tool from Microsoft that enables users to connect to a **wide range of data sources**, **create interactive visualizations**, and share insights through reports and dashboards.
2. It offers a user-friendly interface for data **transformation, modeling, and analysis, leveraging powerful features like DAX (Data Analysis Expressions)** and Power Query for advanced calculations and data cleaning.
3. Power BI supports **real-time data updates, collaboration, and secure sharing across teams,** making it an essential platform for organizations to **analyze, visualize,** and make **data-driven decisions**. With its integration into the Microsoft ecosystem and cloud capabilities, Power BI ensures scalability, security, and accessibility for users on desktops, mobile devices, and web platforms.

Power BI Dashboard



Total Sales
51.93M

Total Products
295

Total Products Sold
67K

Price-of-Discoun t
4.98M

Total Cities
269
Count of City

Max orders from single customer
1102

Due-Date
13-06-2...

Max-online -order
2560

Related Technical Documentation

1. “Empowering Manufacturing Industries with Industrial Data Lake solution on Azure cloud “, 2024, Technical Project Report by Maheswaran V , https://www.dataeverconsulting.com/whitepapers

2. “Data Engineering Solution for an Inventory management Database on SAP-HANA”, 2024 , technical Project Report by Maheswaran V , https://www.dataeverconsulting.com/whitepapers

Conclusions

1. Given a big industrial dataset spread across different Storage repository in different types and different formats, We built a Big data lake on Azure cloud.
2. Using the Azure data services and using the Extract, Transform, Load and business analytics process, convert the raw data into a clean data and extracted insights from the data and presented the valuable data in the form of a Dashboard.

Bigdata Analytics on Azure cloud