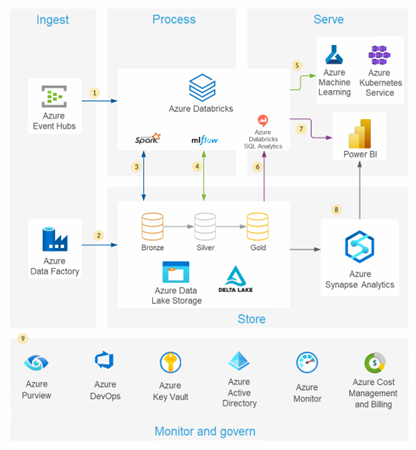
**Project Solution Document: Formula 1 Data Pipeline on Azure**

# 1. Project Overview

The Formula 1 Data Pipeline project is designed to ingest, process, and analyze Formula 1 racing data using Azure services and Databricks. This project enables efficient data processing, storage, and querying capabilities to support analytics and reporting.

# 2. Solution Architecture



The architecture comprises the following components:

## 2.1 Azure Services

* Azure Resource Group: Centralized management of all Azure resources.
* Azure Storage Account with ADLS Gen2: Stores raw, processed, and presentation data.
* Azure Databricks Workspace: Executes data ingestion and transformation.
* Azure Key Vault: Securely stores secrets like Databricks credentials.
* Azure IAM Roles: Manages access permissions for Databricks to access ADLS.

## 2.2 Data Flow

1. Data Ingestion:

* Raw data (CSV, JSON) is uploaded to Azure Data Lake Gen2 (Raw Layer).
* Databricks notebooks extract data from the raw layer and perform transformations.

1. Data Processing:

* Spark-based transformations clean and prepare data.
* Partitioning and Delta Lake formats optimize query performance.

1. Data Storage:

* Processed data is stored in ADLS (Processed Layer) for efficient querying.
* Aggregated and analytical datasets are stored in the Presentation Layer.

1. Data Consumption:

* Querying via Databricks SQL and visualization tools.

# 3. Infrastructure Deployment

Infrastructure is deployed using Terraform to ensure scalability and repeatability.

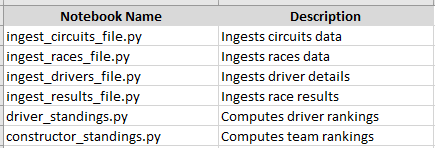
## 3.1 Terraform Modules

* Storage Account & ADLS Gen2: Manages data lake.
* Databricks Workspace: Deploys a Spark-based analytics environment.
* Key Vault: Stores secrets securely.
* IAM Roles: Grants necessary permissions.

# 4. Security Considerations

* Role-Based Access Control (RBAC): Databricks has restricted access based on IAM roles.
* Secure Storage: Secrets are stored in Azure Key Vault.
* Data Encryption: ADLS Gen2 encrypts data at rest and in transit.

# 5. Data Processing Notebooks



# 6. Deployment Steps

1. Set Up Azure Environment:

* Deploy resources using Terraform.

1. Configure Databricks:

* Mount ADLS containers.
* Store credentials in Key Vault.

1. Execute Notebooks:

* Run data ingestion scripts.
* Perform data transformations.

1. Validate Data:

* Query results to ensure correctness.

1. Integrate with Reporting Tools:

* Use Power BI or other tools to visualize data.

# 7. Monitoring & Maintenance

* Databricks Job Scheduling: Automate pipeline execution.
* Logging & Alerts: Monitor execution failures and data inconsistencies.
* Performance Optimization: Tune Spark jobs and storage configurations.

# 8. Conclusion

This project enables scalable, efficient, and secure processing of Formula 1 racing data using Azure cloud technologies and Databricks, ensuring optimized performance for analytics and reporting.