ScholarNest

Introduction to Data Engineering

Data Engineers – Reference Architecture

ScholarNest

Image Source: Google Cloud Documentation

Business data owners

What do they do?

- Develop and Manage Operational Systems
  - Banking Apps
  - E-Commerce Apps
  - OTT Applications
  - IoT Applications

Data engineering team

What do they do?

- Collect data
- Transform
  - Quality Check
  - Standardize
  - Prepare/Model
- Facilitate Consumption

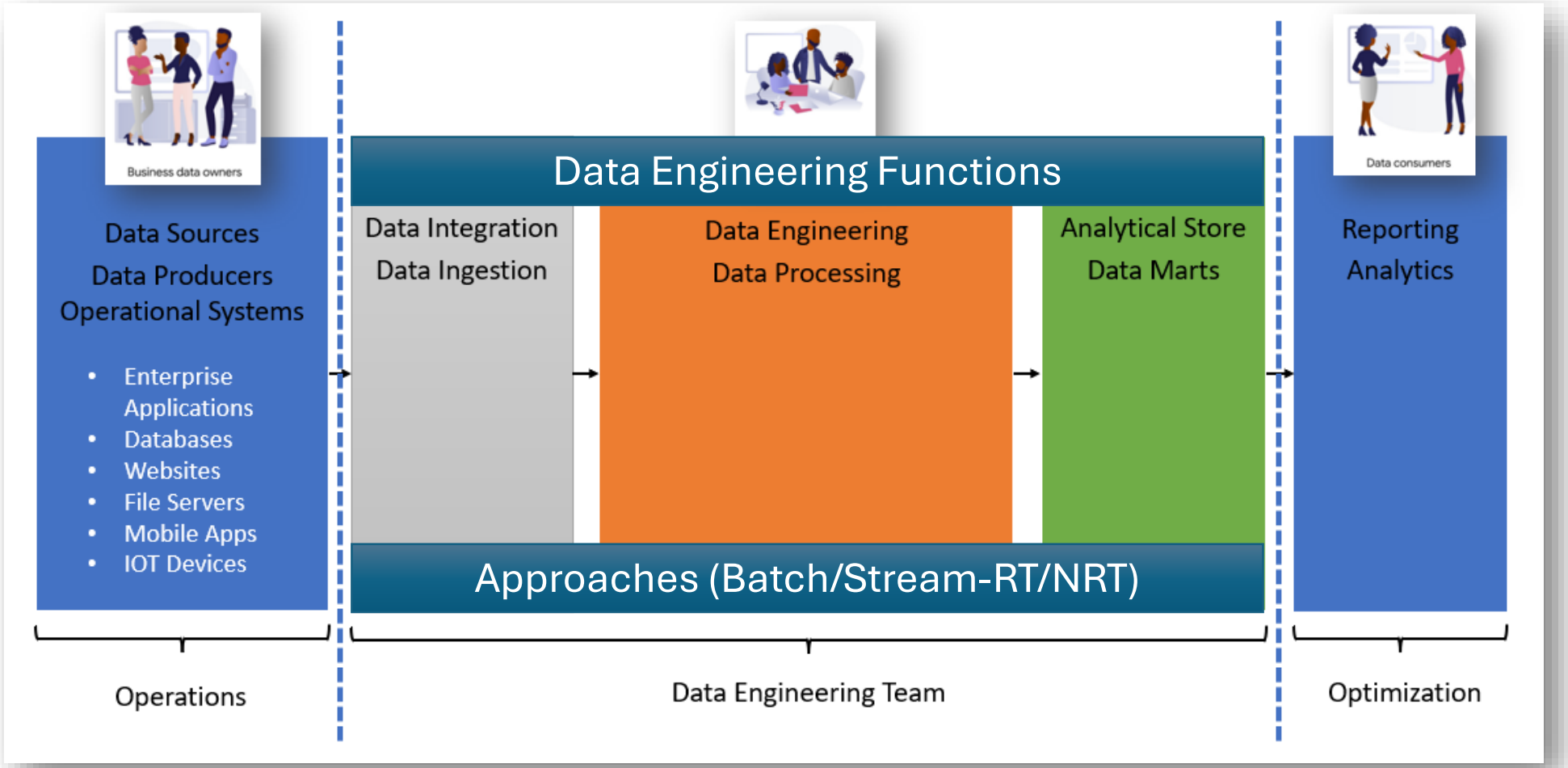
Data consumers

What do they do?

- Optimize
  - Fraud Prevention
- Grow
  - Recommendations
- Monitor/Report
  - Sales/Revenue

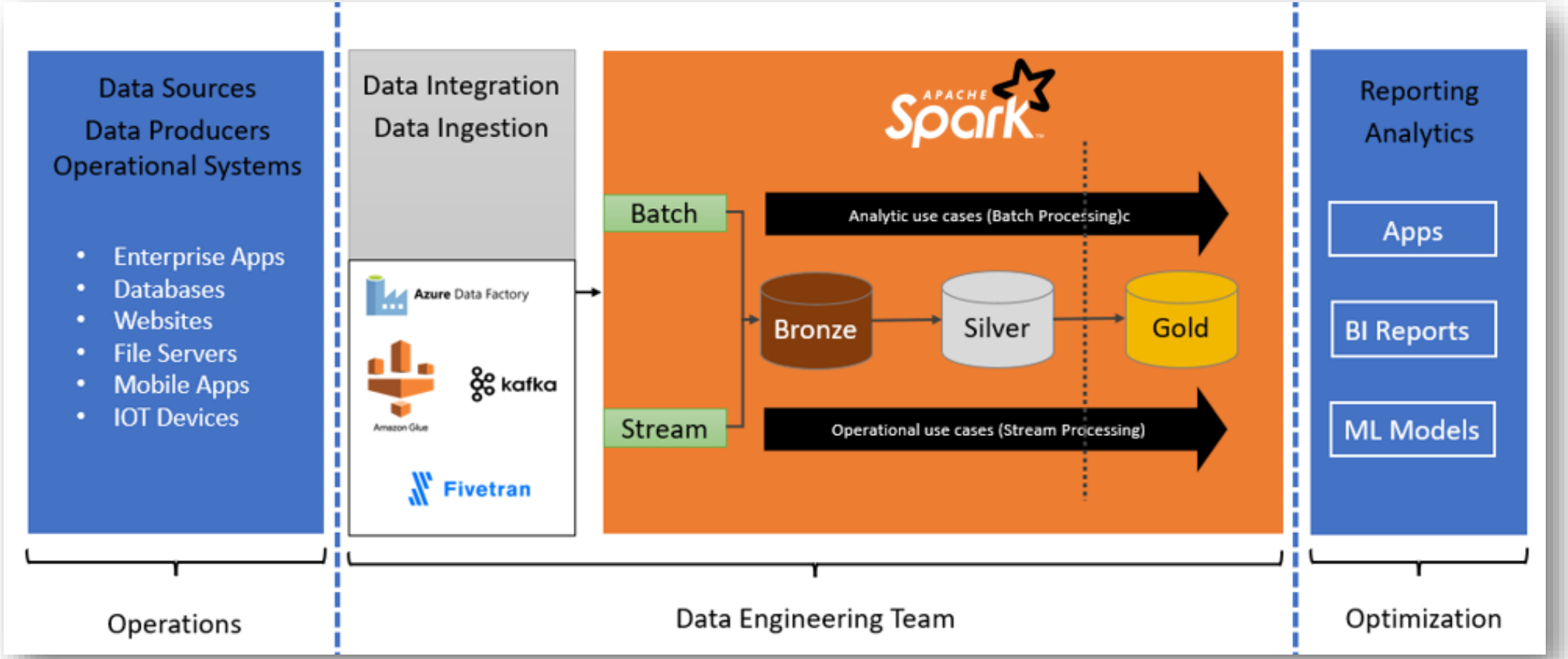
Data Engineering Platform – Reference Architecture

ScholarNest



Lakehouse Medallion Architecture

ScholarNest



ScholarNest

Introduction to Apache Spark

ScholarNest

What is Apache Spark?

Apache Spark™ is an engine for executing data engineering, stream processing, and machine learning on distributed clusters.

Capabilities

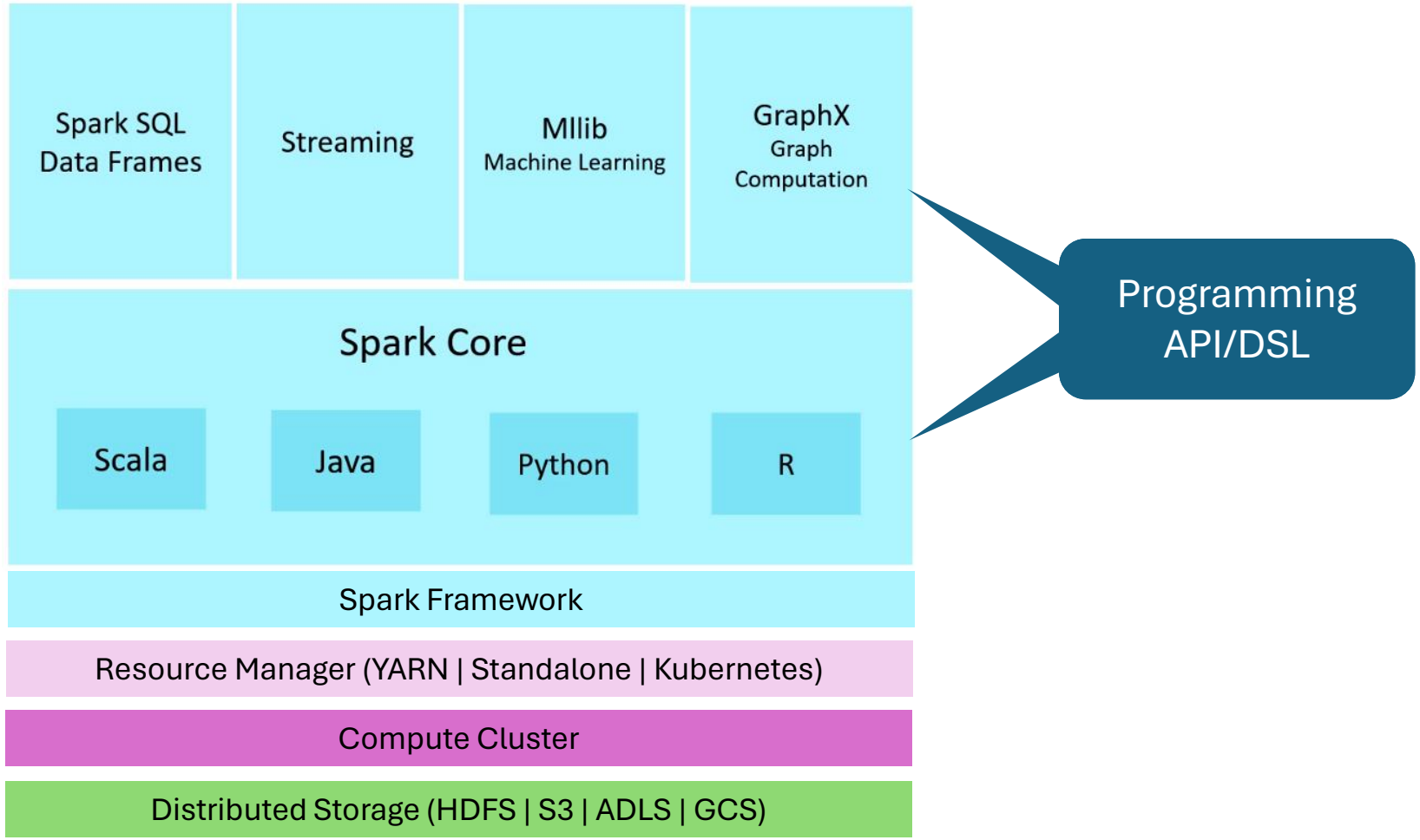
- ANSI SQL
- Batch Processing API
- Stream Processing API
- Graph Processing API
- Machine Learning API

Who is using Apache Spark?

Thousands of companies, including 80% of the Fortune 500, use Apache Spark™.

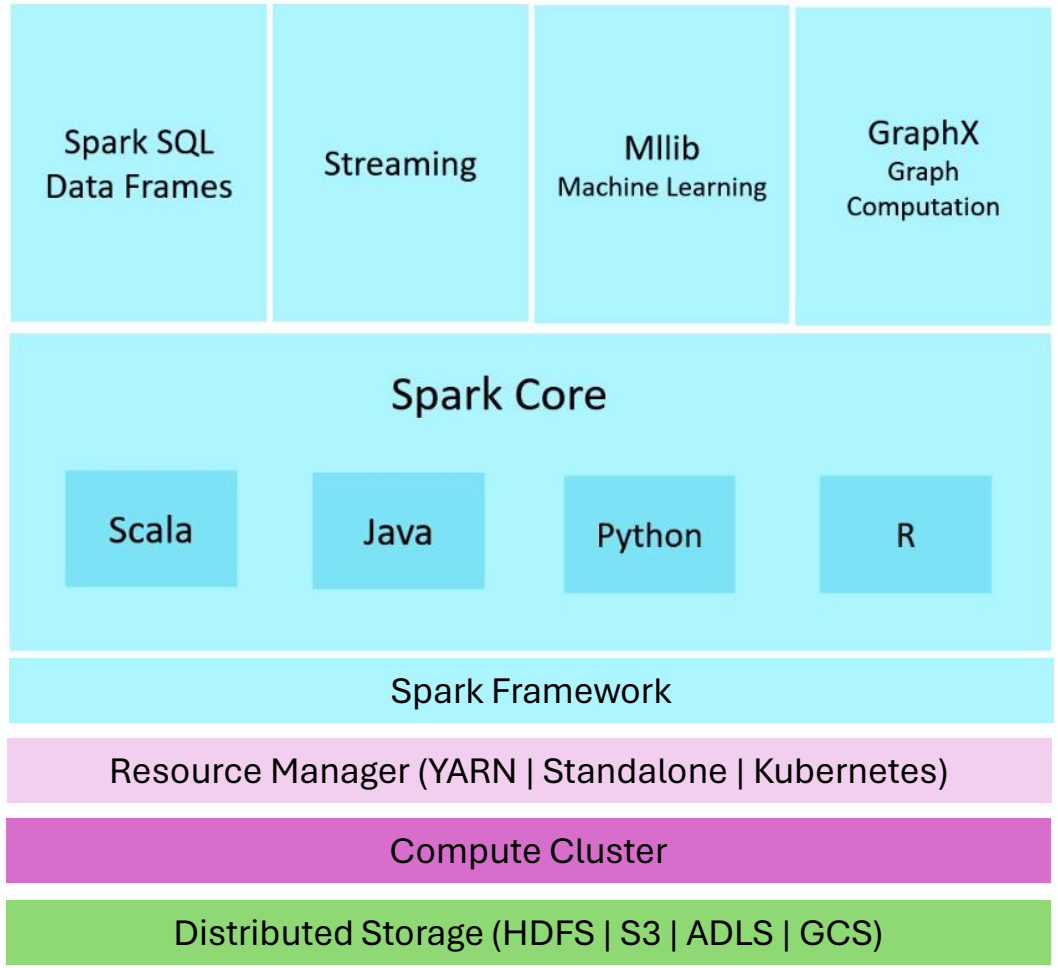
What is Apache Spark – A Unified Framework

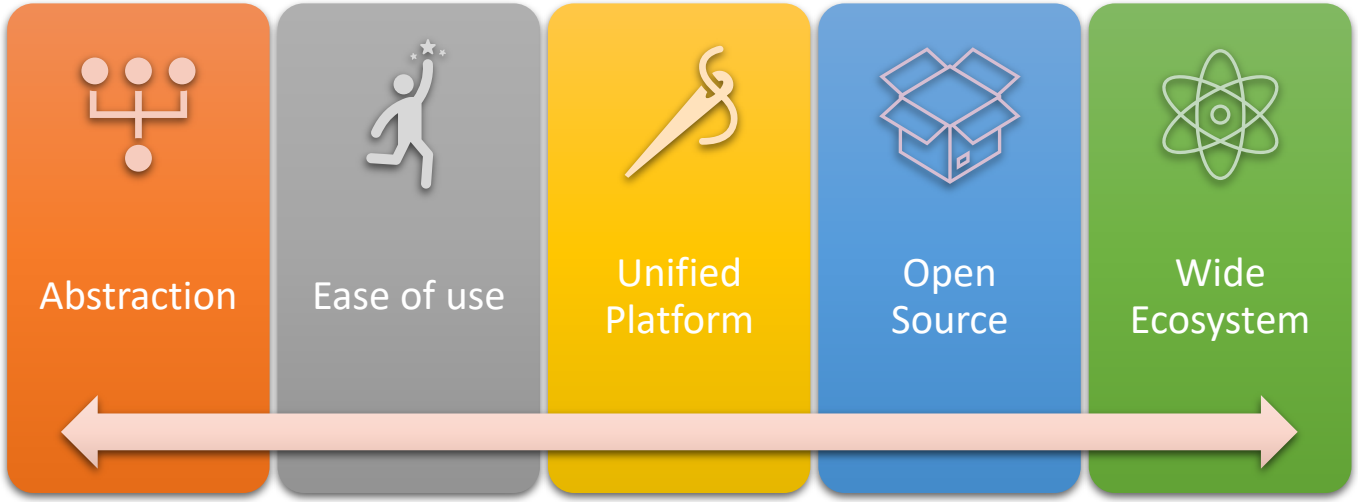
ScholarNest



Why Apache Spark?

ScholarNest





Missing features from Apache Spark

ScholarNest

- Data Storage Infrastructure
- ACID Transaction capabilities
- Metadata Catalog
- Cluster Management
- Automation APIs and Tools

Spark Platforms

ScholarNest

- Cloudera Hadoop Platform
- Amazon EMR
- Azure HDInsight
- Google Data Proc
- Databricks Platform

ScholarNest

Introduction to Databricks

Databricks Features

ScholarNest

- Spark as Cloud-Native Technology
- Secure Cloud Storage Integration
- ACID Transaction via Delta Lake Integration
- Unity Catalog for Metadata Management
- Cluster Management
- Photon Query Engine
- Notebooks and Workspace
- Administration Controls
- Optimized Spark Runtime
- Automation Tools

Databricks Cloud

Databricks Cloud – Key Integrations

ScholarNest

Service	Azure	AWS	GCP
CI/CD	Azure DevOps, GitHub Enterprise	AWS Code Build, AWS Code Deploy, AWS Code Pipeline	Google Cloud Build, Google Cloud Deploy
Data warehouse	Azure Synapse Analytics	Amazon Redshift	BigQuery
Data Integration	Azure Data Factory	AWS Glue, Amazon Data Pipeline	Google Cloud Data Fusion
Messaging	Azure Service Bus, Azure Event Hubs	AWS Kinesis, Amazon SNS, Amazon SQS	Google Pub/Sub
Workflow orchestration	Azure Data Factory	Amazon Data Pipeline, AWS Glue, Apache Airflow	Cloud Composer
Document data	Azure Cosmos DB	Amazon DocumentDB	Firestore
NoSQL - Key/Value	Azure Cosmos DB	Amazon DynamoDB	Cloud Bigtable
RDBMS	Azure SQL Database	Amazon Aurora, Amazon RDS	Cloud SQL
Storage Transfer	Azure Data Factory, Azure Storage Mover	AWS Storage Gateway, AWS Data Sync	Storage Transfer Service
Network connectivity	Azure Virtual Private Network	AWS Virtual Private Network	Cloud VPN
Audit logging	Azure Audit Logs	AWS CloudTrail	Cloud Audit Logs
Key management	Azure Key Vault	AWS KMS	Cloud KMS
Identity	Azure Identity Management	AWS IAM	Google Cloud IAM
Storage	Azure Blob Storage - ADLS Gen2	Amazon S3	Google Cloud Storage