DHL Group : Building an Enterprise Grade Cloud Unified Data Lakehouse Platform on Azure

Hitesh Sahni

Head of Cloud Data Platforms and Solutions,

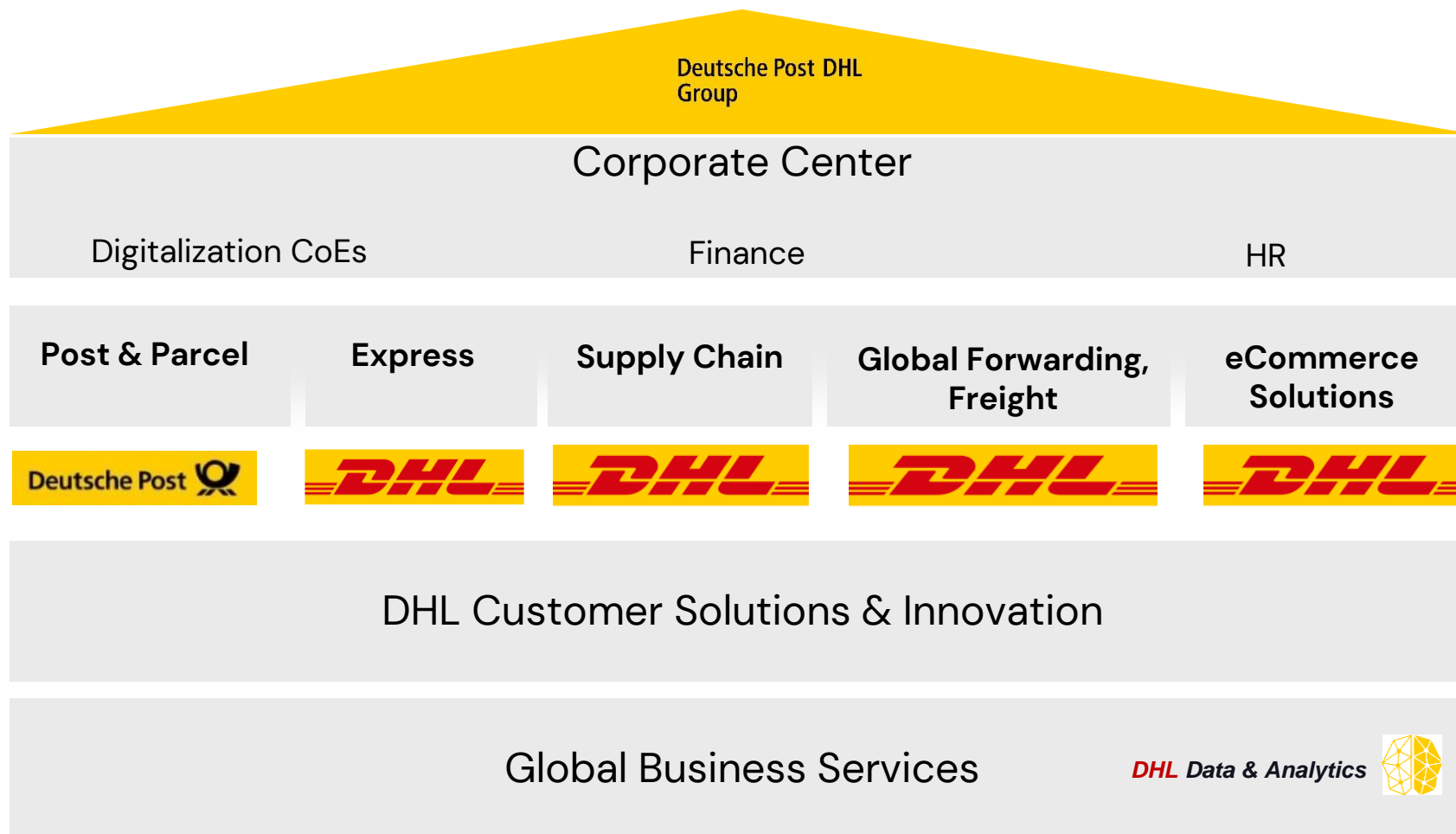DHL Data & Analytics, DHL Group

AGENDA

CLICK TO EDIT

DHL Group overview

Europe
351,428 Employees

Americas
92,753 Employees

Asia/Pacific
84,036 Employees

Other regions
19,241 Employees

~57mn letters &
~ 5m parcels per workday
in Germany

~ 995,000 international express
shipments per day

~3.8m tons of **air freight**
~3.2m TEU of **ocean freight**

~13.2m square meters of
warehouse space in
contract logistics

NEED FOR DIVISIONAL AND XDIV CLOUD DATA PLATFORMS

Key Challenges

- Large Enterprise with many different divisions and functions with their own business focus and capabilities.
    - Different BUs/Divisions have varied IT & Data Analytics landscape.
    - Different BU's and teams are at different level of maturity and adoption of data analytics journey.
    - Different BU's have different data sources and various types of analytics usecases.

- BU's and functions large enough need more or full control, ownership and flexibility choice for their data platforms and technologies, processes, etc.

- BUs existing DWH systems are either not scalable or too costly to scale and maintain to meet future data management and analytics needs.

- BU's existing DBs and DWH systems have limited or no capabilities for Big Data workloads like unstructured data, Streaming analytics, advanced analytics (ML, AI).

- Very huge effort and time required for integration of new data into analytics platform.

DIGITILIZATION IS FUELED BY DATA AT DHL Group

KNOW YOUR DATA

Knowing the data content and meaning, the teams who create and consume your data is the first pre-requisite to create business insights.
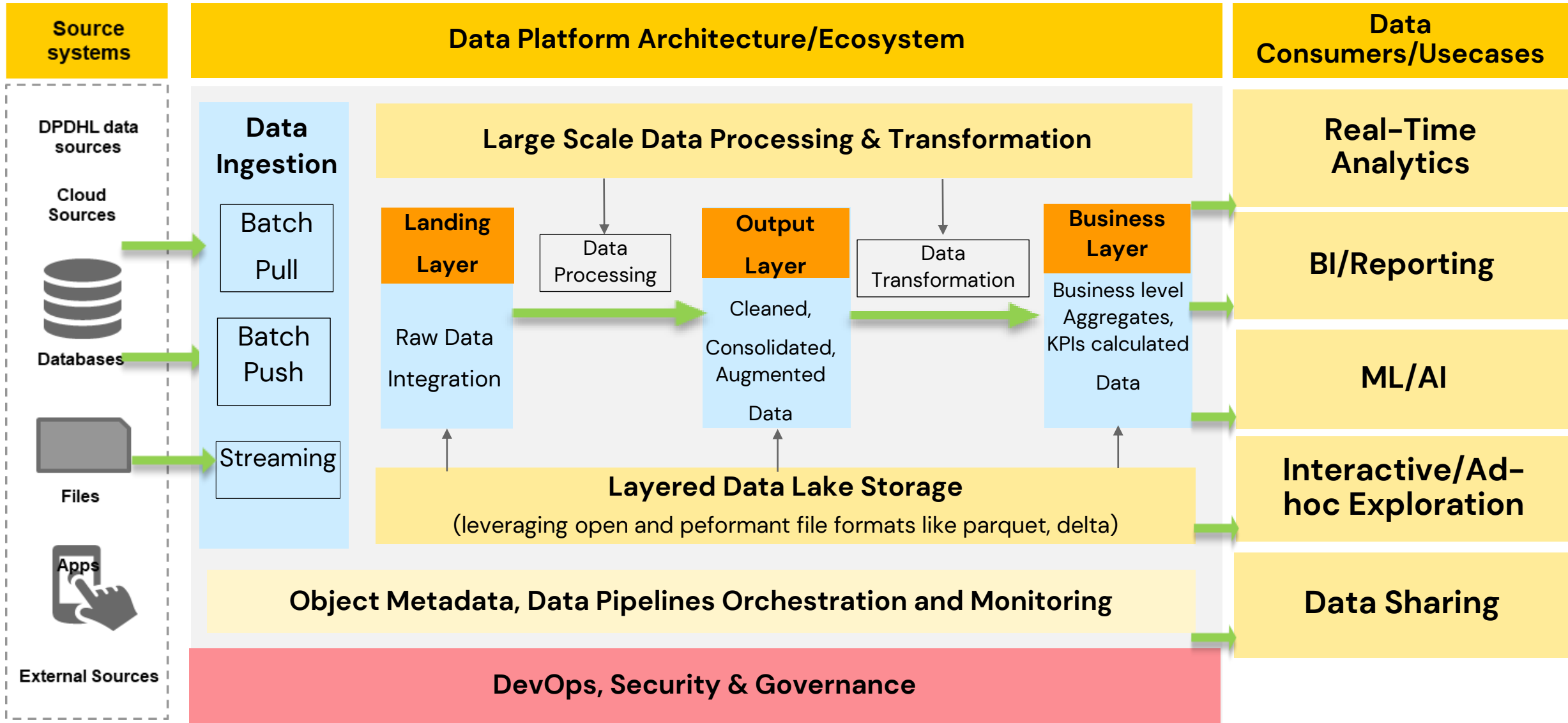
OWN YOUR DATA

Taking ownership is essential to continuously improve data quality and make it in a compliant way.

USE YOUR DATA

Using your data to understand, improve, and auotmate our business processes is key to ultimately create business value.

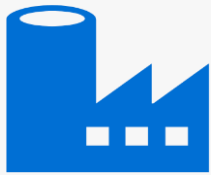# STANDARDIZED CLOUD DATA PLATFORM ARCHITECTURE AT DHL GROUP

Data Platform Technology Components and their usage

Central Metadata/Config DB Model

Data Ingestion	**Data Processing**	**Data Preparations/ Transformations**	**Query Layer**
Azure Data Factory	Azure Databricks Spark Jobs	Azure Databricks Spark SQL Jobs	Azure Synapse **Serverless** **Data Lake query Engine**
Highly Available SHIR Cluster (on VMs)	Separated Job Clusters for each source system data loads	Separated Job Clusters for each source system data loads	Databricks SQL (in evaluation)

Layered Data Lake Storage
1 Storage account per source system

ADLS Gen2 storage accounts

DELTA LAKE   Parquet

End to End Data Pipelines Orchestration and Monitoring Capabilities
(extended with custom orchestration capabilities – smart scheduling, customized scheduling, multi–file event triggers)

Azure Data Factory

DATA LAKE STORAGE DESIGN PRINCIPLES & SETUP EXAMPLE

Data Provisioning

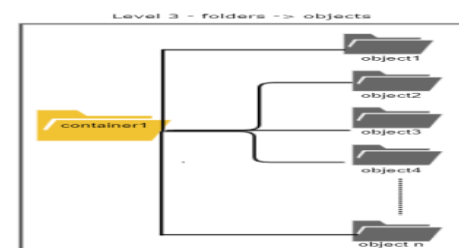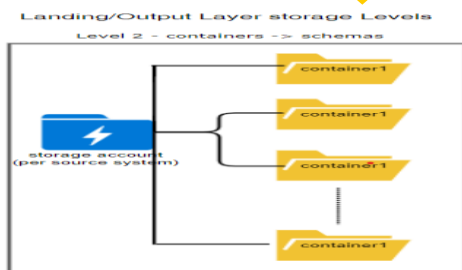### Landing Layer (Ingestion)

- Raw data from source systems
  - (XML, csv files, parquet, avro, etc.)

- Data from source systems stored 1:1 (no data manipulation)

- Not accessible for end users

- Only for data ingestion

Output Layer (Data Standardization)

- Data Validation and standardization for consumption:
  - Flattening, Schema Validation, Parquet conversion, Consolidation, Historization, Anonymization, etc.

- Same structure as Landing layer

- Addition of data annotations if needed.

Usecase Layer (Business Logic)

- Usecases get their own space for storing their processed data and applications.

- Usecases should get data needed from outbound (or landing) layer.

- No data from Source systems should directly flow into Usecase space.

Level 1 - storage accounts -> src systems
storage account (per source system)
storage account (per source system)
storage account (per source system)
storage account (per source system)

Landing/Output Layer storage Levels
Level 2 - containers -> schemas
storage account (per source system)
container1
container1
container1
container1

Level 3 - folders -> objects
container1
object1
object2
object3
object4
object n

Domain Driven Design (or Source system in case domains don't exist)

HOW WE BUILT THE CLOUD DATA PLATFORM
KEY TENETS

Metadata Driven Framework.

- Flexible and customizable config DB Model.
- New objects onboarding only requires a config addition

Reusable Data Pipeline Templates

- Pre-built and tested ingestion and processing pipelines.
- Different load patterns available.

Leverage Out of the Box connectors

- for disparate systems (SFTP, SFDC, API, RDBMS (oracle, Teradata), etc.)
- Tested and Adapted to fit the working environment.

Workload Isolation

- Separate Job clusters for different source systems data processing.
- Job clusters of different capacities for usecase/team workloads per needs.

Automation

- Reusable IAC deployment for ADF,
- Automated Deployments and Release management

Security & DevOps Baked-in

- Security by design in all elements.
- Code Version control, CICD
- Release management.

Multi-Tenant Platform Architecture and Setup Approach

ExpressRoute connectivity establishing secure channel for data transfer between DHL sources and Azure Data Platform.
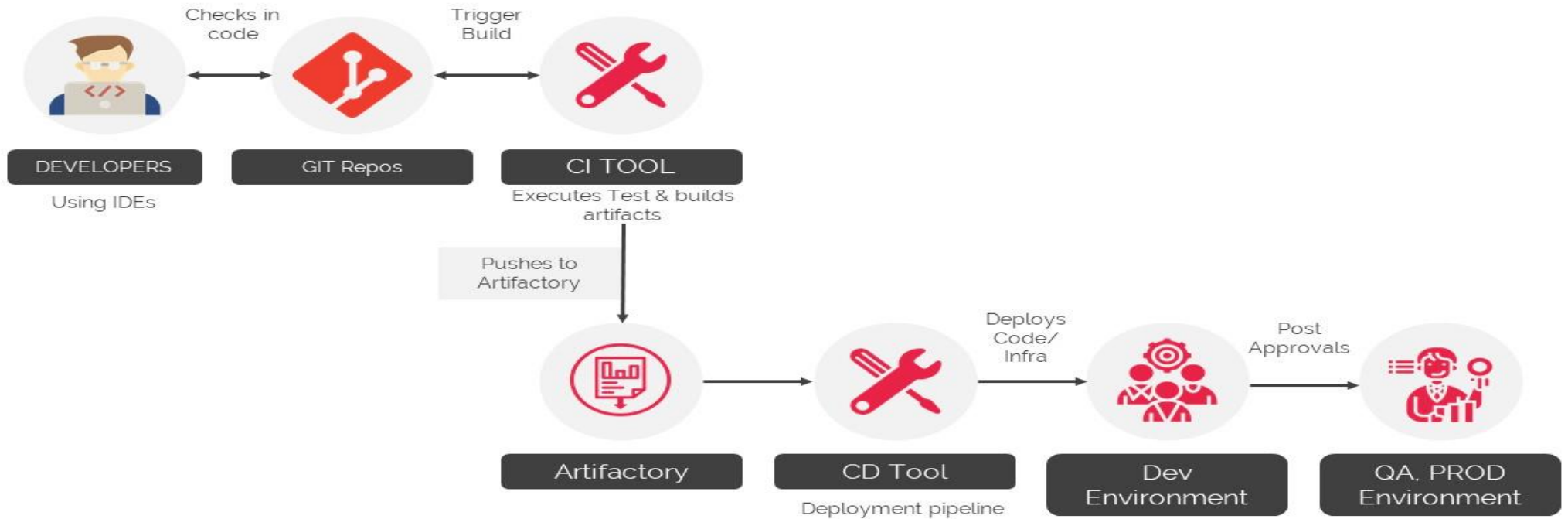
Concept of **Core Platform + Use cases area.**

- **Core Platform** provides reusable components and data foundation.

- Use cases/teams get their own segregated space to operate independently.

IAC and automated builds for repeatable and version-controlled platform deployments.
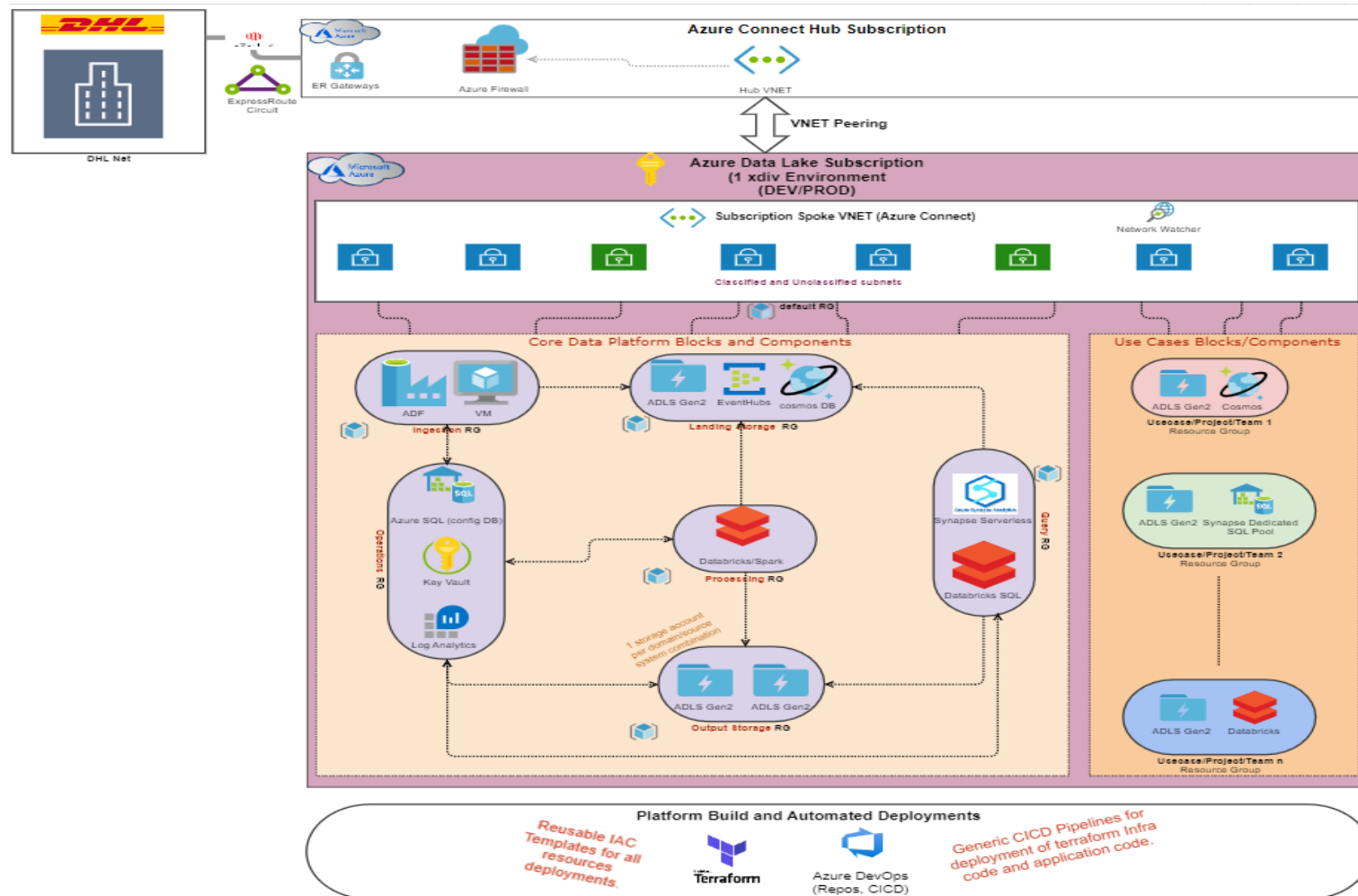
Automated Deployment and Release Management driven by Azure DevOps

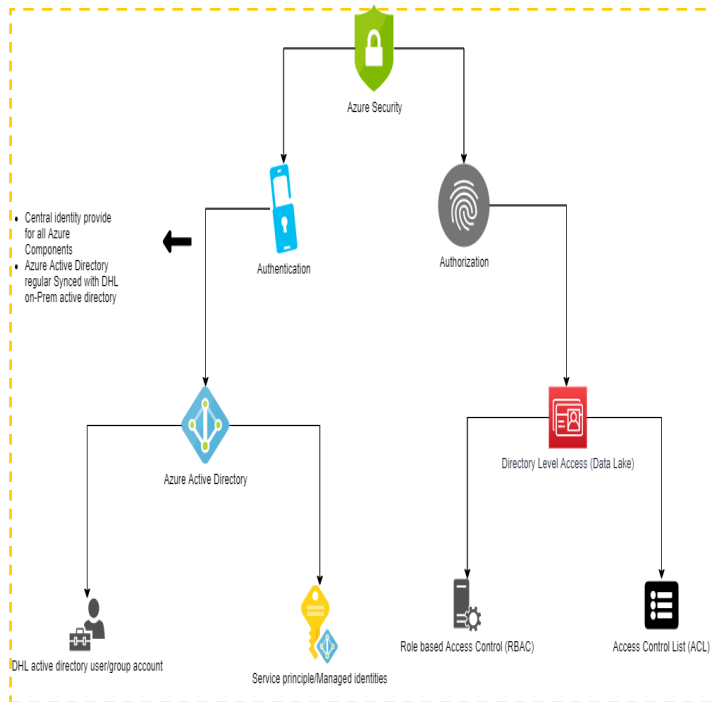# Security & Governance as first Class Citizen

PRIVATE CONNECTIVITY	AUTHENTICATION & AUTHORIZATION BUILT–IN across all Azure resources of platform	DEVSECOPS CHECKS & CONTROLS	PLATFORM GOVERNANCE

PRIVATE CONNECTIVITY

- Dedicated Network Line via EXPRESSROUTE
- Private Link & Endpoints connectivity
- Secure DevOps Integration
- Databricks with VNET injection

AUTHENTICATION & AUTHORIZATION BUILT–IN across all Azure resources of platform

Azure Security

- Central identity provide for all Azure Components
- Azure Active Directory regular Synced with DHL on-Prem active directory

Authentication — Authorization

Azure Active Directory — Directory Level Access (Data Lake)

DHL active directory user/group account — Service principle/Managed identities — Role based Access Control (RBAC) — Access Control List (ACL)

DEVSECOPS CHECKS & CONTROLS

- Infra drift checks & security scans (tfsec)
- Cloud Security Posture Management via Prisma
- Security and Vulnerability scans and management (Qualys)
- Static code scans and security scans via Sonarqube

PLATFORM GOVERNANCE

- AZURE POLICIES
- RESOURCE TAGGING
- COST ALERTS

DHL

Cloud Data Lakehouse Platforms established as Foundational Next-Gen Analytics platforms at DHL Group

State-of-the-art Cloud based data & analytics services

Large Scale Big Data Storage & Compute for Data processing

Enabling all types of analytics workloads from Single data platform

Mature enterprise setup

BU and X-Div Azure Data Platforms

DSC Azure Data Lake Platform	Microsoft Azure	eCS US Azure Data Platform	Microsoft Azure
eCS Global Azure Data Platform	Microsoft Azure	PnP Azure Data platform (BI 2025)	Microsoft Azure
ITS Data platform	Microsoft Azure	X-Div Azure Data Platform	Microsoft Azure

Automated and Repeatable Platform Build & Deployment Blueprints

Global Promixity

Integration Platform and Data Security and Governance

Maximizing Business Value across the DHL group with Unified Cloud Data Lakehouse Platforms and Data Management Frameworks

Usecases on cloud data platforms generating impact in all areas of our Logistics business

DSC
Smart Volume prediction, Transport Analytics, Warehouse Data Integration

GBS
X-div. Customer Carbon Reporting

DHL eC
Operational reports, e.g. delay reports, Customer At Risk, Volume forecasting

FIN
Accounts receivable and payable reporting

CSI
Sales Analytics, e.g. white spot analysis based on x-div data

GBS
Corporate Real Estate Reporting

THANK YOU

SESSION FEEDBACK

Session Title: Strategie und Technologie: Advanced Analytics mit Azure (DHL Group)

https://aka.ms/AzSum-S017