WEEK 3
DAY – 1

4/09/2023

Nabiha Khan

WORKING WITH DATA STORAGE

Storage

- 2 types
- Simple storage
- ADLS – Azure Datalake Storage
- Benefits of using Azure to store data:
- Automated Backup – mitigates the risk of losing data in any circumstance
- Global Replication – to protect data against any planned/unplanned event
- Encryption Capabilities – transmitted data encryption, azure key vault
- Multiple Data Types
- Support for Data Analytics
- Storage Tiers
- Virtual Disks

Storage

- Comparing Azure to on-premises storage:

- Cost effectiveness – pay as you go pricing model
- Reliability – data backup, load balancing, disaster recovery, and data replication
- Storage Types – azure provides multiple store options, providing the best streaming
- Agility – flexibility to create new services in minutes

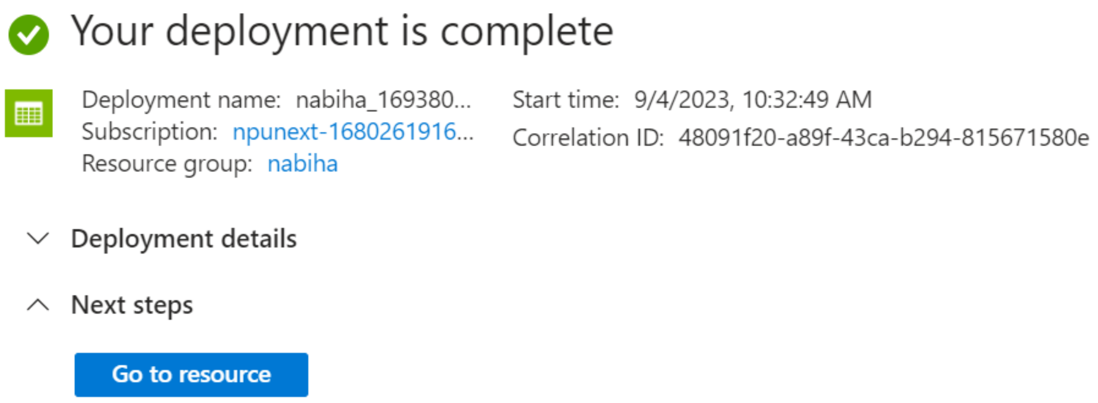
Create Azure Storage Account

- Storage Accounts
- Container that groups a set of storage services
- Data Diversity – Cost Sensitivity – Management Overhead
- Storage Account Settings:
- Subscription
- Location
- Performance
- Version
- Access Tier
- Replication

Create Azure Storage Account

- Creation Tools

- Simple/complex



Your deployment is complete

Deployment name: nabiha_169380...
Subscription: npunext-1680261916...
Resource group: nabiha

Start time: 9/4/2023, 10:32:49 AM
Correlation ID: 48091f20-a89f-43ca-b294-815671580e

AZURE DATA LAKE STORAGE

Azure Data Lake Storage – Gen II

- Big Data Hadoop Access
- Security
- Performance
- Redundancy

Azure blob Storage vs Data Lake Gen – II

- Azure Blob – Flat namespace
- Data Lake Gen-II – Hierarchical namespace

Processing Big Data with Azure Data Lake Store

1. Ingestion
2. Store
3. Prep & Train
4. Model & Serve

Big Data Use Cases

- Modern Data Warehouse
- Advanced Analytics (cosmos DB comes into play)
- Real – Time Analytics (data comes from dynamic resources)

DAY – 2

05/09/2023

AZURE DATA FACTORY

Orchestrating data movement with Azure Data Factory

E – Extract Data from source
T – Transformation
L – Load to destination

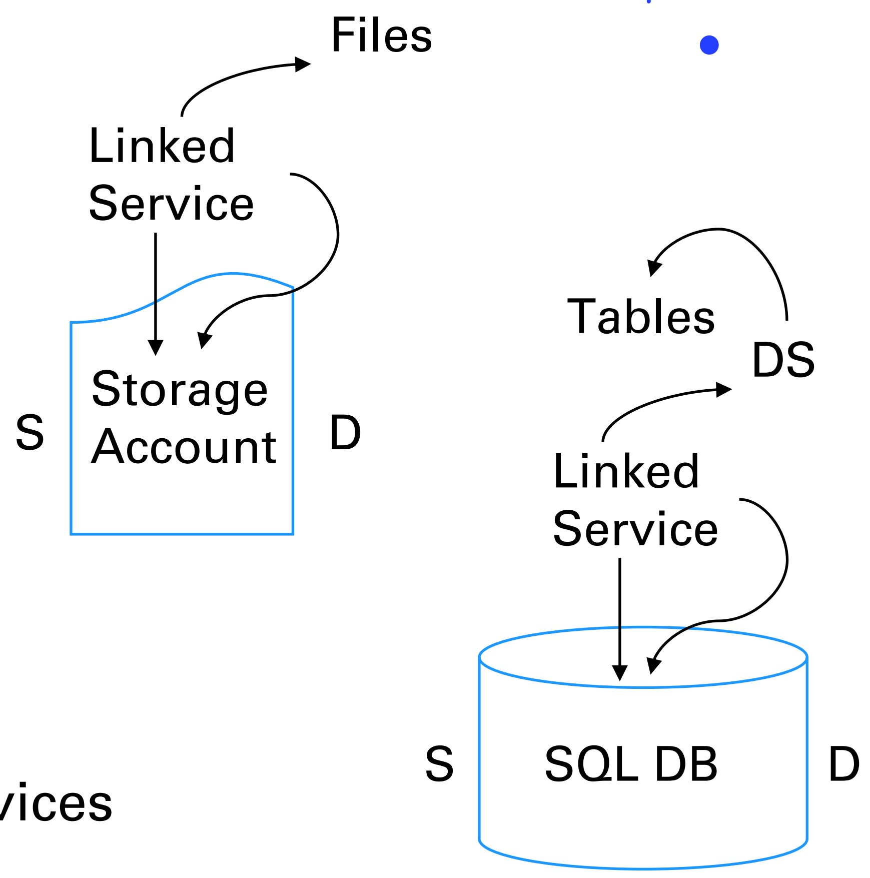
E – Extract Data
L – Load into its raw format
T – Transformation (on demand)

Orchestrating data movement with Azure Data Factory

- Azure Data Factory:
- A brick through which data modification takes place through in the cloud
- Automate the resources
- Analyze
- The Data Factory Process
- Connect & Collect
    - Ingest
    - Prepare
- Transform & Enrich
- Publish
- Monitor

Orchestrating data movement with Azure Data Factory

- Azure Data Factory Components
- Linked Service
- Data Set
- Activity
- Pipeline
- Control Flow – perform
- Parameters
- Integration Runtime – acts as a bridge between two services



Orchestrating data movement with Azure Data Factory

- Azure Data Factory Security

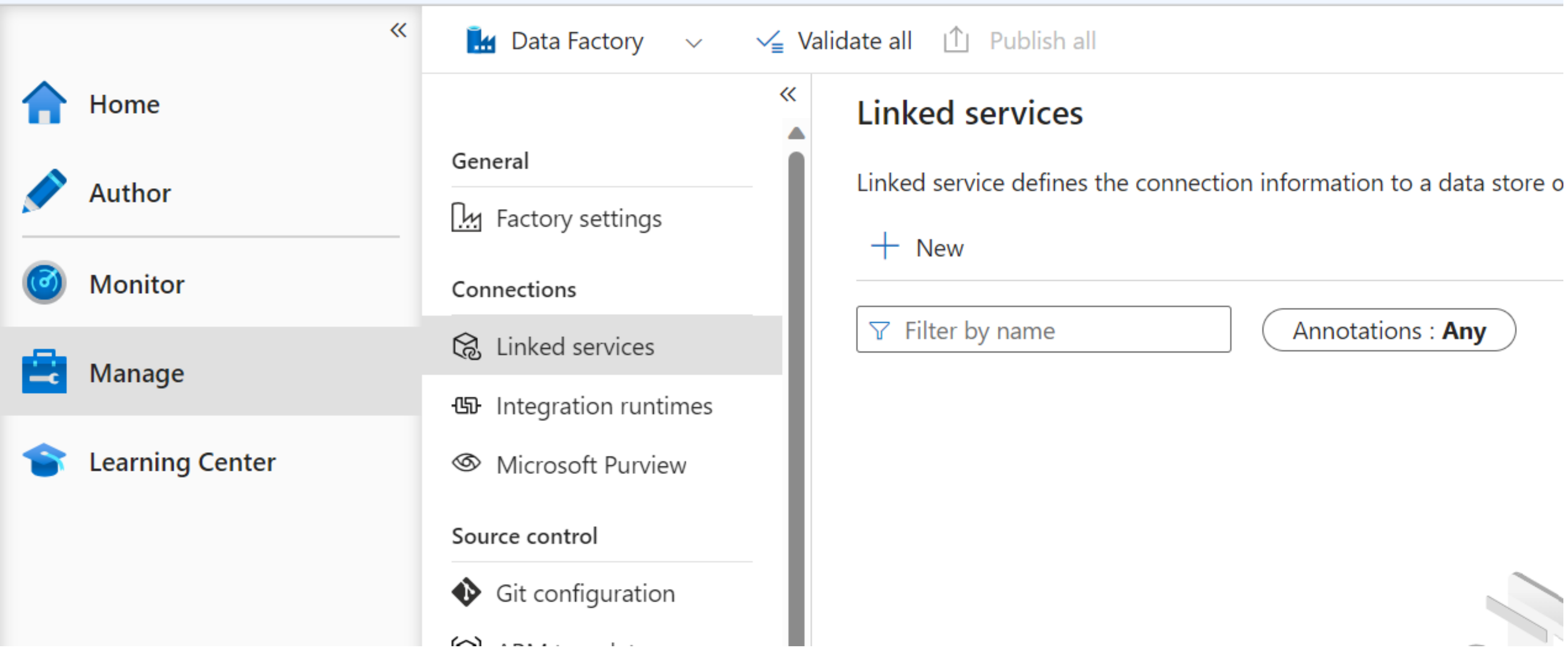
Data Factory Contributor Rule

- Azure Data Factory Components

- Linked Service – Data Lake Store & Databricks
- DataSet – Background (DataSet name, properties, structure, availability, policy)
- Activity – 40 (data movement, data transformation, & control activities)
- Pipeline – when trigger goes data travels from data source. Grouping of logically related activities, scheduling, managed & monitored
- Control Flow
- Integration Runtime – 2 types -> AutoResolve IR (cloud-to-cloud)-> Self hosted IR (on-premise-to-cloud)
- Parameters

Ingesting & Transforming Data

- Creating Azure Data Factory



Linked services

Linked service defines the connection information to a data store o

DAY – 3

06/09/2023

Ingesting & Transforming Data

- Ingesting data with the copy activity

- Lab
- LookUp

1. Read the information
2. Filter

DAY – 4

07/09/2023

Ingesting & Transforming Data

- Get MetaData
- Filter
- If Else
- For Each

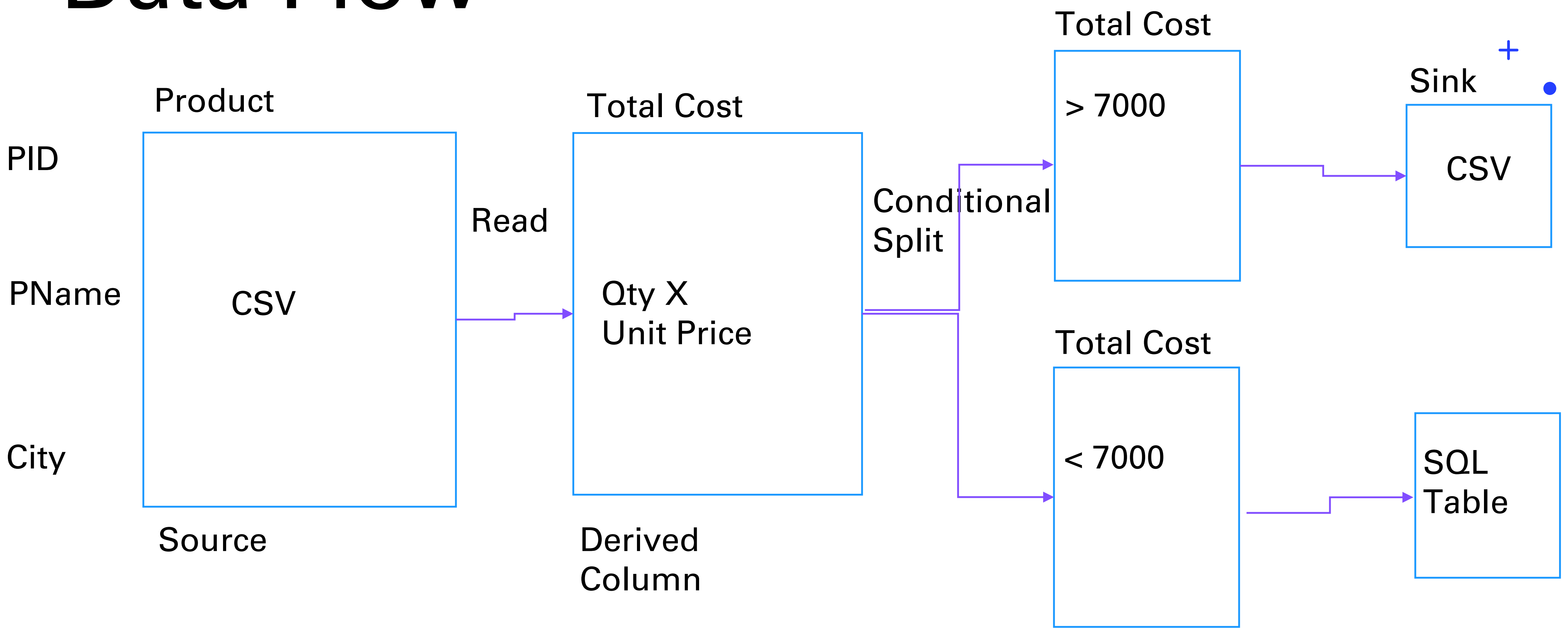
Data Flow

- Source Activity
- Sink Transformation/Activity
- Union Transformation
- Surrogate Key Transformation
- Conditional Split Transformation
- Derived Column Transformation

- Concepts

- Mapping
- Validation of resources

Data Flow


Product
PID
PName
City
CSV
Source
Read
Total Cost
Qty X
Unit Price
Derived
Column
Conditional
Split
Total Cost
> 7000
Total Cost
< 7000
Sink
CSV
SQL
Table

DAY – 5

08/09/2023

Data Flow

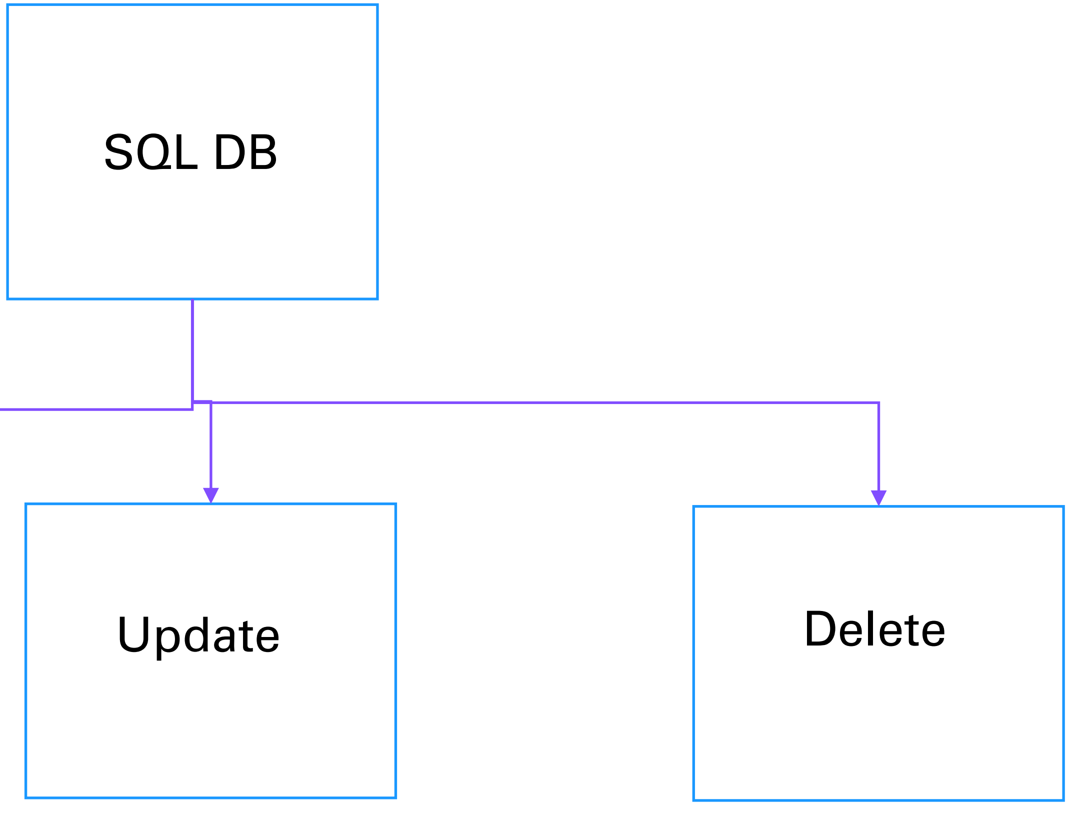
- Lab – Conditional Split

API Integration

- Microservices & Application
- Amazon
- Walmart
- Allows application to build & use business logics, data & visualization forms as a service
- Demo

CDC

- Change Data Capture
- Incremental Update



Data Flow

- The CDC control task and data flow components

- Initial Extraction
- Incremental Extraction

Monitoring & Troubleshooting

- General Azure Monitoring Capabilities

-> Azure Monitor

- Metric Data – Threshold value, CPU utilization, data transaction ie visual appearance
- Log Data – transactions or logs capturing
- Alerts – Triggers information

-> Monitoring The Network

- Network Performance Monitor
- Application Gateway Analysis

-> Diagnose & Solve Problem

Monitoring & Troubleshooting

- Troubleshooting the common data issues

-> Connectivity Issues

- Unable to connect to the data platform
- Authentication failures
- Cosmos DB Mongo DB API errors
- SQL database failover

-> Performance Issues

- Data Lake Storage
- SQL Database
- Cosmos DB
- Colocation of resources
- SQL Data Warehouse