Activity 01: Data Lake Design & Security Considerations

Wide World Importers is ready to build a pipeline that copies their sales transactions from a table in an Oracle database to the data lake.

Requirements

- The pipeline that copies data will run on a scheduled basis, once per day.
- They would like to ingest this raw data applying the minimal amount of transformations to it.
- They want to ensure that their data lake always contains a copy of the original data, so that if their downstream processing has calculation or transformation errors, they can always re-compute from the original.
- Additionally, they want to avoid the file format prescribing what tools can be used to examine and process the data by making sure that the file format selected can be used by the broadest possible range of of industry standard tools.
- The folder structure needs to be performant for typical exploratory and analytic queries for this type of data.

Example of the data

WWI has provided the following example of the data to you. You can assume they have hand selected rows that are *most* representative of the data.

SaleKey	CityKey	CustomerKey	BillToCustomerKey	StockItemKey	DeliveryDateKey	SalespersonKey	WWIInvoiceID	Description	Packag
294018	98706	0	0	25	2012-01-04	156	57894	Black and orange, handle with care despatch tape 48mmx75m	Each
294019	98706	0	0	216	2012-01-04	156	57894	USB, food flash drive - sushi roll	Each
294020	98706	0	0	168	2012-01-04	156	57894	IT joke mug - keyboard not found press F1 to continue (White)	Each
294021	98706	0	0	100	2012-01-04	156	57894	Dinosaur battery-powered slippers (Green) L	Each

Whiteboard

Open your whiteboard for the event, and in the area for Activity 1 provide your answers to the following challenges.

The following challenges are already present within the whiteboard template provided.

Challenges

1. What file format should they use for the raw data? Why did you recommend this file format, provide at least two reasons?
2. What specific settings should WWI use in configuring the way the dataset is serialized to disk (pay particular attention to the `Description` field? Why did you suggest these?
3. Diagram the folder structure you would recommend they use in the hierarchical filesystem. Be sure to indicate filesystem, folders and files and describe how each layer (filesystem, folder and file) derives its name.
4. How does your folder structure support query performance for typical exploratory and analytic queries for this type of data?
5. WWI consider this data confidential, because if it were to fall into the hands of the competition it would cause irreparable harm. Diagram how you would deploy the data lake and secure access to the data lake endpoint? Be sure to illustrate how data flows between your Azure Synapse Analytics Workspace and the data lake and explain why this addresses WWI's requirements. Use the icons provided in the palette to diagram your solution.
6. WWI wants to enforce that any kind of modifications to sales data can happen in the current year only, while allowing all authorized users to query the entirety of data. Regarding the folder structure you previously recommended to WWI, how would accomplish this using RBAC and ACLs? Explain what actions would need to be taken at the start and end of the year 2020. Diagram your security groups, built-in roles and access permissions using the provided palette.