**Day 15 Notes**

**Name : Podutur Lahari - DE126**

**Date : 28-11-2024**

### **Delta Table with Existing Data in Databricks**

**Overview of Delta Lake:**Delta Lake is a powerful open-source layer that adds reliability and functionality to traditional data lakes. It supports ACID transactions to maintain data integrity, scalable management of metadata for large datasets, and combines both real-time streaming and batch processing. This makes it suitable for handling complex and dynamic data pipelines.

**Steps to Create a Delta Table in Azure Databricks:**

1. Uploading Data to Databricks File System (DBFS):  
   The first step is to upload the data you want to work with into the Databricks environment.
   * Use the Databricks platform to create a new folder for storing files.
   * Data files can be uploaded directly from a local system into this folder.
   * The uploaded files are stored in paths that act as references for later processing.
2. Writing Data in Delta Format:  
   Once the data is uploaded, it needs to be written in Delta format.
   * This involves reading the original data from its location in DBFS and saving it in another folder designated for Delta-formatted data.
   * Delta format ensures that data is stored in a structured, reliable way, adding metadata for tracking changes and providing a transaction log.
3. Creating a Delta Table:  
   After preparing the data in Delta format, a Delta table can be created.
   * A Delta table is essentially a structured representation of the data, allowing users to query it easily.
   * The process creates a schema (the organization of data), and the data itself remains in its Delta-formatted location. This separation means the table can be dropped without losing the actual data.
4. Visualizing the Delta Table Data:  
   The final step involves querying and visualizing the data from the Delta table.
   * This allows users to confirm the structure and accuracy of the data and use it for further analysis.
   * Queries return tabular results, which can be explored for insights.

### **Azure Databricks Visualization**

Creating Visualizations in Databricks:  
Visualization is a core feature in Databricks, enabling users to better understand and interpret data through charts and graphs. Here's how you create and manage visualizations:

1. Building a Visualization:
   * Start with a table containing data.
   * Use the “+” symbol in the interface to add a new visualization.
   * Select a chart type that suits your data, such as scatter plots, bar charts, or line graphs.
2. Customization Options:
   * Change the colors, labels, and layout of your visualization to match the required aesthetics and clarity.
   * These changes help in highlighting specific trends or comparisons within the data.
3. Editing and Managing Visualizations:
   * Once a visualization is created, you can modify it for better presentation or clarity.
   * The interface allows for deleting, downloading, or viewing these visualizations as needed.

**Data Profiling:**Data profiling is an additional feature in Databricks that provides a statistical summary of the dataset.

* This process examines trends, patterns, and anomalies in the data.
* It also supports ordering and filtering operations, enabling deeper analysis of subsets of data.

**Key Benefits:**

* Visualizations help in understanding data trends and relationships quickly.
* Data profiling aids in data quality assessment, ensuring that only clean and accurate data is used for decision-making.

**Key Features of Delta Lake:**

1. ACID Transactions:
   * Guarantees reliable and consistent data updates even in high-concurrency environments.
   * Ensures atomicity, consistency, isolation, and durability, critical for robust data pipelines.
2. Unified Batch and Streaming Data Processing:
   * Integrates seamlessly with both real-time streaming data and traditional batch processing, eliminating the need for separate systems.
3. Scalable Metadata Management:
   * Efficiently handles metadata for massive datasets, enabling faster queries and updates.
4. Version Control and Data Lineage:
   * Maintains a log of changes (known as delta\_log), providing a history of all operations performed on the data.
   * Facilitates rollbacks and reproducibility by allowing users to access previous versions of the data.

**Applications of Delta Lake:**

* Suitable for real-time data analytics, such as in IoT systems or stock market applications, where incoming data must be processed alongside historical data.
* Ideal for building ETL pipelines, where data needs to be transformed and analyzed reliably before loading into downstream systems.

Creating Delta Tables in Databricks:  
Delta tables represent a structured way of interacting with data stored in Delta format. They act as a bridge between raw data files and analytical tools, allowing users to query the data efficiently. This combination of raw storage and structured querying provides flexibility in how data is stored and accessed.

**Importance of Delta Tables:**

* Efficiency: Provides faster read and write operations compared to traditional formats like CSV or Parquet.
* Data Quality: Delta tables ensure the consistency and integrity of the data, making them suitable for critical applications.
* Flexibility: The data remains in a Delta format on the file system, so users can access it through multiple tools or frameworks beyond Databricks.

### **Azure Databricks Visualization**

Significance of Data Visualization:  
Data visualization is the process of representing data in graphical formats such as charts, graphs, and dashboards. It allows users to explore trends, patterns, and anomalies visually, which is often more intuitive than interpreting raw data or text-based summaries.

**Why Use Databricks for Visualization?**

Databricks integrates visualization capabilities directly into its notebooks, enabling users to quickly transition from querying data to analyzing it visually. This integration streamlines workflows by eliminating the need to use separate tools for data analysis and visualization.

**Types of Visualizations Supported:**

1. Scatter Plots: Useful for identifying relationships and correlations between variables.
2. Bar Charts: Ideal for comparing categories or discrete groups.
3. Line Charts: Effective for tracking changes over time or trends in data.
4. Pie Charts: Best for visualizing proportions or percentages.

**Benefits of Visualization in Azure Databricks:**

1. Real-Time Insights: Visualizations update dynamically as the underlying data changes, providing real-time insights.
2. Collaboration: Visuals created in Databricks can be shared with team members, fostering better decision-making.
3. Customization: Users can edit and modify visualizations to focus on specific data attributes or aesthetics.